OY LEUANGTHONG and
CLAYTON V. DEUTSCH, Editors

Geostatistics Banff 2004

Banff 2004
Seventh International
Geostatistics Congress

QUANTITATIVE GEOLOGY AND GEOSTATISTICS

Springer

GEOSTATISTICS BANFF 2004

Volume 1

Quantitative Geology and Geostatistics

VOLUME 14/1

The titles published in this series are listed at the end of this volume.

GEOSTATISTICS BANFF 2004

Volume 1

Edited by

OY LEUANGTHONG

University of Alberta,
Edmonton, Canada

and

CLAYTON V. DEUTSCH

University of Alberta,
Edmonton, Canada

Springer

A C.I.P. Catalogue record for this book is available from the Library of Congress.

ISBN-10 1-4020-3515-2 (HB) (indivisible set)
ISBN-13 978-1-4020-3515-9 (HB) (indivisible set)
ISBN-10 1-4020-3610-8 (e-book)
ISBN-13 978-1-4020-3610-1 (e-book)

Published by Springer,
P.O. Box 17, 3300 AA Dordrecht, The Netherlands.

www.springeronline.com

Printed on acid-free paper

Printed in the Netherlands.

TABLE OF CONTENTS

VOLUME 1

PLENARY

MINING

VOLUME 2

PETROLEUM

ENVIRONMENTAL

THEORY & SELECTED TOPICS

FOREWORD

The return of the congress to North America after 20 years of absence could not have been in a more ideal location. The beauty of Banff and the many offerings of the Rocky Mountains was the perfect background for a week of interesting and innovative discussions on the past, present and future of geostatistics.

The congress was well attended with approximately 200 delegates from 19 countries across six continents. There was a broad spectrum of students and seasoned geostatisticians who shared their knowledge in many areas of study including mining, petroleum, and environmental applications. You will find 119 papers in this two volume set. All papers were presented at the congress and have been peer-reviewed. They are grouped by the different sessions that were held in Banff and are in the order of presentation.

These papers provide a permanent record of different theoretical perspectives from the last four years. Not all of these ideas will stand the test of time and practice; however, their originality will endure. The practical applications in these proceedings provide nuggets of wisdom to those struggling to apply geostatistics in the best possible way. Students and practitioners will be digging through these papers for many years to come.

Oy Leuangthong
Clayton V. Deutsch

ACKNOWLEDGMENTS

We would like to thank the industry sponsors who contributed generously to the overall success and quality of the congress:

De Beers Canada
Earth Decision Sciences
Maptek Chile Ltda.
Mira Geoscience
Nexen Inc.
Petro-Canada
Placer Dome Inc.
Statios LLC
Total

We would also like to thank all the chair people, reviewers, volunteers, contributors and delegates who helped to make the congress a success. Special thanks to R. Mohan Srivastava for his help with the artwork. We thank the International Geostatistics committee for their comments and suggestions:

Winfred Assibey-Bonsu (South Africa)
Jef Caers (USA)
Clayton Deutsch (Canada)
Roussos Dimitrakopoulos (Australia)
Peter Dowd (Australia)
J. Jaime Gómez-Hernández (Spain)
Christian Lantuéjoul (France)

The staff and students of the Centre for Computational Geostatistics at the University of Alberta gave generously of their time to all aspects of organizing the congress. Amanda Potts (research administrator) provided essential administrative support. Sandra Correa, Paula Larrondo, Jason McLennan, Chad Neufeld, Weishan Ren, Stefan Zanon, and Linan Zhang (graduate students) provided invaluable assistance before and during the congress.

LIST OF PARTICIPANTS

Abrahamsen, Petter, Norwegian Computing Center, Po Box 114 Blindern, Gaustadalleen 23, OSLO, NO0314, NORWAY

Alapetite, Julien, Earth Decision Science, 3 Route De Grenoble, MOIRANS, FR38430, FRANCE

Almeida, Jose, CIGA, Faculdade Ciencias Tecnologia, Universidade Nova Lisboa, MONTE DA CAPARICA, PORTUGAL

Arpat, Burc, Stanford University, Department of Petroleum Engineering, STANFORD, CA 94305, USA

Assibey- Bonsu, Winfred, Gold Fields Mining Services Ltd, 24 St. Andrews Rd, ZA 2193, SOUTH AFRICA

Assis Carlos, Luis, Anglo American Brasil Ltda, 502- Setor Santa Genoveva, Av. Interlandia, GOIANIA 74672-360, BRAZIL

Aug, Christophe, Centre de Geostatistique, 35 Rue Saint- Honore, Ecole Des Mines De Paris, FONTAINEBLEAU F77305, FRANCE

Bankes, Paul, Teck Cominco Limited, 600-200 Burrard Street, VANCOUVER, BC V6C 3L9, CANADA

Barbour, Russell, EPH/Yale School of Medicine, 60 College St. Rm 600, P. O. Box 208034, NEW HAVEN, CT 06520, USA

Barrera, Alvaro, University of Texas at Austin, 4900 E Oltorf St Apt 538, AUSTIN, TX 78741, USA

Bellentani, Giuseppe, ENI E&P, Via Emilia 1-5 Pal., Uff.-5013 E, SAN DONATO MILANESE, 20097, ITALY

Berckmans, Arne, NIRAS, Kunstlaan 14, BRUSSSELS, 1210, BELGUIM

Bernard- Michel, Caroline, Ecole des Mines de Paris, 35 Rue Saint Honore FONTAINEBLEAU, F77305, FRANCE

Bertoli, Olivier, Quantitative Geoscience Pty Lt , Po Box 1304, FREMANTLE, 6959, AUSTRALIA

Beucher, Helene, Ecole des Mines de Paris, 35 Rue Saint Honore, Centre De Geostatistique, FONTAINEBLEAU, F77305, FRANCE

Biver, Pierre, TOTAL SA, C S T J F, Ba 3112, Avenue Larribau, PAU, 64000, FRANCE

Blackney, Paul, Snowden Mining In. Consultants, P O Box 77, West Parth, 87 Colin St, W A 6872, PERTH, 6005, AUSTRALIA

Blaha, Petr, Holcim Group Support Ltd., Holderbank, HOLDERBANK, 5113, CHILE

Boucher, Alexandre, Stanford University, 151 Calderon Ave, Apt 232, MOUNTAIN VIEW, CA 94041, USA

Bourgault, Gilles, Computer Modelling Group Ltd, 3512 33 Street N W, Office #200, CALGARY, AB T2L 2A6, CANADA

Brega, Fausto, ENI S.P.A- E& P Division, Via Emilia, 1, San Donato Milanese, MILAN, 20097, ITALY

Brown, Gavin, De Beers Consolidated Mines, PO Box 350, CAPE TOWN, 8000, SOUTH AFRICA

Brown, Steven, Nexen Inc, 801-7th Avenue S W, CALGARY, AB T2P 3P7. CANADA

Buecker, Christian, R W E Dea A G, Ueberseering 40, HAMBURG, 22297, GERMANY

Bush, David, De Beers, P Bag X01 Southdale, Cnr. Crownwood Rd & Diamond Dr, JOHANNESBURG, ZA2135, SOUTH AFRICA

Buxton, Bruce, Battelle Memorial Institute, 505 King Avenue, COLUMBUS, OH, 43201, USA

Caers, Jef, Stanford University, Petroleum Engineering, 367 Panama St, STANFORD, CA, 94305, USA

Carvalho, Jorge, Dep. Minas, Faculty Of Engineering, PORTO, 4200-465, PORTUGAL

Castro, Scarlet, Stanford University, 361 Green Earth Sciences Bldg., 367 Panama Street, STANFORD, CA, 94305, USA

Caumon, Guillaume, Stanford University, 367 Panama St., Petroleum Engineering Dept, STANFORD, CA, 94305, USA

Chappell, Adrian, University of Salford, Environment & Life Sciences, Peel Building, The Crescent, SALFORD, M5 4WT, UK

Chelak, Robert, Roxar Canada, 1200 815-8 Th Ave SW, CALGARY, AB, T2P 3P2, CANADA

Chen, Zhuoheng, Geological Survey of Canada, 3303-33rd Street N.W., CALGARY, AB, T2L 2A7, CANADA

Cheng, Qiuming, York University, 4700 Keele Street, NORTH YORK, ON, M3J 1P3, CANADA

Chiles, Jean- Paul, Ecole des Mines de Paris, 35 Rue Saint Honore, Centre De Geostatistique, FONTAINEBLEAU, FR77305, FRANCE

Coburn, Timothy, Abilene Christian University, Acu Box 29315, ABILENE, TX, 79699, USA

Cornah, Alastair, University of Exeter, Camborne School Of Mines, Trevenson Rd, CORNWALL, TR15 3SE, UK

Correa Montero, Sandra, University of Alberta, Department of Civil & Environmental. Engineering, 3-133 Markin/CNRL Natural Resources Engineering Facility, EDMONTON, AB, T6G 2W2, CANADA

Costa, Joao Felipe, UFRGS, Av. Osvaldo Aranha 99/504, PORTO ALEGRE, 90035190, BRAZIL

Costa, Marcelo, BRAZIL

Da Silva, Emidio, University A. Neto, Faculdade De Engenharis-d De, Eng. Minas Av. 21 De Janeiro, LUANDA, 1756, ANGOLA

Dagbert, Michel, Geostat Systems Int. Inc, 10 Blvd Seigneurie E, Suite 203, BLAINVILLE, QC, J7C 3V5, CANADA

Dahle, Pal, Norwegian Computing Center, P O Box 114, Blindern, Gaustadalleen 23, OSLO, NO0314, NORWAY

Daly, Colin, Roxar Ltd, Pinnacle House, 17-25 Hartfield Rd, WIMBLEDON, SW19 3SE, UK

De Visser, Jan, RSG Global

Della Rossa, Ernesto, ENI S.P.A, Eni E& P- Apsi, Via Emilia 1, San Donato Milanese, MILANO, 20097, ITALY

Deutsch, Clayton, University of Alberta, Department. of Civil & Environmental Engineering, 3-133 Markin/CNRL Natural Resources Engineering Facility, EDMONTON, AB, T6G 2W2, CANADA

Dimitrakopoulos, Roussos, University of Queensland, Wh Bryan Mining Geo. Res. Cen, Richards Building, BRISBANE, 4072, AUSTRALIA

Dohm, Christina, Anglo Operations Ltd., P O Box 61587, MARSHALLTOWN, 2107, JOHANNESBURG, 2000, SOUTH AFRICA

Dose, Thies, RWE Dea AG, Ueberseering 40, Hamburg, HAMBURG, D22297, GERMANY

Dowd, Peter, Faculty of Engineering, Computer and Mathematical Sciences, University of Adelaide, ADELAIDE, SA 5005, AUSTRALIA

Dube, Pascal, Cameco, 243 Beckett Green, SASKATOON, SK S7N 4W1, CANADA

Duggan, Sean, De Beers Consolidated Mines, 117 Hertzog Boulevard, CAPE TOWN, 8001, SOUTH AFRICA

Dungan, Jennifer, NASA Ames Research Center

Eidsvik, Jo, Statoil, Ark Ebbels V 10, Statoil Research Center, TRONDHEIM, NO7032, NORWAY

Emery, Xavier , University of Chile, Dept. Of Mining Engineering, Avenida Tupper 2069, SANTIAGO, 8320000, CHILE

Faechner, Ty, Assiniboine Community College, 1430 Victoria Ave East, BRANDON, MB, R7A 2A9, CANADA

Fisher, Thomas, Colorado School of Mines, 1211-6th Street, GOLDEN, CO, 80403, USA

Francois Bongarcon, Dominique, Agoratek International, 345 Stanford Center Pmb# 432, PALO ALTO, CA, 94304, USA

Froidevaux, Roland, FSS Consultants SA, 9, Rue Boissonnas, GENEVA, 1256, SWITZERLAND

Frykman, Peter, Geological Suvery of Denmark, G E U S, Oster Voldgade 10, COPENHAGEN, 1350, DENMARK

Gabriel, Edith, INRA, Domaine St Paul, Site Agropic, AVIGNON, 84914, FRANCE

Gallagher Jr., Joseph, ConocoPhillips, Reservoir Sciences, 269 Geoscience Bldg, BARTLESVILLE, OK, 74004, USA

Garcia, Michel, FSS International, 1956, Avenue Roger Salengro, CHAVILLE, FR92370, FRANCE

Garner, David, Conoco Phillips Canada Ltd., P.o. Box 130, 401- 9th Ave S W, CALGARY, AB, T2P 2H7, CANADA

Glacken, Ian, Snowden Mining In. Consultants, P O Box 77 West Perth, 87 Colin St, WA 6872, PERTH, 6005, AUSTRALIA

Gloaguen, Erwan, Ecole Polytechnique, C. P. 6079 Succ. Centre-ville, MONTREAL, QC, H3C 3A7, CANADA

Gomez- Hernandez, Jaime, Univ. Politecnica de Valencia, Escuela De Ing. De Caminos, Camino De Vera S/n, VALENCIA, EP46022, SPAIN

Gonzalez, Eric, Maptek South America, 5 Norte 112, Vina Del Mar, VINA DEL MAR, 2520180, CHILE

Goovaerts, Pierre, BioMedware, 516 N State St, ANN ARBOR, MI, 48104, USA

Gray, James, Teck Cominco Ltd, 600-200 Burrard St, VANCOUVER, BC, V6C 3L9, CANADA

Grills, John Andrew, De Beers Consolidated Mines, 117 Hertzog Boulevard, PO Box 350, CAPE TOWN, 8000, SOUTH AFRICA

Gringarten, Emmanuel, Earth Decision Sciences, 11011 Richmond Ave, Suite 350, HOUSTON, TX, 77042, USA

Guenard, Cindy, Talisman Energy Inc., Suite 3400, 888-3rd Street S.W., CALGARY, AB, T2P 5C5, CANADA

Guibal, Daniel, SRK Consulting, P. O. Box 943 West Perth, 1064 Hay Street, WEST PERTH, 6005, AUSTRALIA

Harding, Andrew, Chevron Texaco, 6001 Bollinger Canyon Road, DANVILLE, CA, 94506, USA

Hauge, Ragnar, Norwegian Computing Center, P. O. Box 114 Blindern, OSLO, 0314, NORWAY

Hayes, Sean, Talisman Energy Inc., Suite 3400, 888-3rd Street S W, CALGARY, AB, T2P 5C5, CANADA

Henry, Emmanuel, AMEC/Ecole Polytechnique, De Montreal, Suite 700, 2020 Winston Park Dr, OAKVILLE, ON, L6H 6X7, CANADA

Hlavka, Christine, NASA, NASA/ AMES Res. Center, Mail Stop 242-4, MOFFETT FIELD, CA, 94035, USA

Hoffman, Todd, Stanford University, 367 Panama Street, Green Earth Sciences Bldg., STANFORD, CA 94305, USA

Hu, Lin Ying, I F P, 1 4 A. De Bois Preau, RUEIL- MALMAISON, 92870, FRANCE

Huysmans, Marijke, University of Leuven, Redingenstraat 16, LEUVEN, 3000, BELGIUM

Isaaks, Edward, Isaaks & Co., 1042 Wilmington Way, REDWOOD CITY, CA, 94062, USA

Jackson, Scott, Quantitative Geoscience Pty Lt, P O Box 1304, FREMANTLE, 6959, AUSTRALIA

Jaquet, Olivier, Colenco Power Engineering Ltd, Taefernstrasse 26, BADEN, 5405, CHILE

John, Abraham, University of Texas at Austin, Cpe 2.502, Department of Petroleum & Geosystems, AUSTIN, TX, 78712, USA

Journel, Andre, Stanford University, Petroleum Engineering Department., STANFORD, CA, 94305, USA

Jutras, Marc, Placer Dome Inc, P. O. Box 49330 Bentall Sa., Suite 1600-1055 Dunsmir St., VANCOUVER, BC, V7X 1P1, CANADA

Kashib, Tarun, EnCana Corporation, 150, 9th Avenue S.W, .P O Box 2850, CALGARY, AB, T2P 2S5, CANADA

Keech, Christopher, Placer Dome Inc, P O Box 49330 Bentall Station, 1600-1055 Dunsmuir St, VANCOUVER, BC, V7X 1P1, CANADA

Khan, Dan, 1601, 8708 - 106th Street, EDMONTON, AB, T6E 4J5, CANADA

Kolbjornsen, Odd, Norwegian Computing Center, P.O. Box 114 Blindern, OSLO, 0314 NORWAY

Koppe, Jair, Univ. of Rio Grande do Sul, Av Cavalhada 5205 Casa 32, PORTO ALEGRE, 91.751-831, BRASIL

Krige, Daniel G, Po Box 121, FLORIDA HILLS, 1716, SOUTH AFRICA

Krishnan, Sunderrajan, Stanford University, #353 Green Earth Sciences, Geological & Environmental Sci, STANFORD, CA, 94305, USA

Kyriakidis, Phaedon, University of California, Dept. of Geography, Ellison Hall 5710, SANTA BARABRA, CA, 93106, USA

Larrondo, Paula, University of Alberta, Department of Civil & Environmental
Engineering, 3-133 Markin/CNRL Natural Resources Engineering Facility,
EDMONTON, AB, T6G 2W2, CANADA

Le Loc`h, Gaelle, Ecole des Mines de Paris, 35 Rue Saint Honore, Centre De
Geostatistique, FONTAINEBLEAU, F77305, FRANCE

Le Ravalec, Mickaela, I F P, 1 4 A. Bois Preau, Rueil-malmaison., 92852, FRANCE

Leuangthong, Oy, University of Alberta, Department of Civil & Environmental
Engineering, 3-133 Markin/CNRL Natural Resources Engineering Facility,
EDMONTON, AB, T6G 2W2, CANADA

Lewis, Richard, Placer Dome Asia Pacific Ltd., G P O Box 465, Brisbane, 4001, 90
Alison Road, RANWICK, AU 2031, AUSTRALIA

Liu, Yuhong, ExxonMobil Upstream Research, P. O. Box 2189, HOUSTON, TX, 77252,
USA

Lodoen, Ole Petter, Norwegian Univ. of Science & T, Dept. Of Mathematical Sciences,
TRONDHEIM, N-7491, NORWAY

Maharaja, Amisha, Stanford University, 51 Dudley Lane, Apt 424, STANFORD, CA,
94305, USA

Marcotte, Denis, Ecole Polytechnique, C. P. 6079 Succ. Centre-ville, SAINT-BRUNO,
QC, J3V 4J7, CANADA

Mattison, Blair, Petro- Canada, 150-6th Ave. SW, CALGARY, AB, T2P 3E3, CANADA

Maunula, Tim, Wardrop Engineering, 2042 Merchants Gate, OAKVILLE, ON, L6M
2Z8, CANADA

Mc Kenna, Sean, Sandia National Laboratories, P O Box 5800 M S 0735,
ALBUQUERQUE, NM, 87185, USA

Mc Lennan, Jason, University of Alberta, Dept. Of Civil & Environmental. Engineering.,
3-133 Markin/CNRL Natural Resources Engineering Facility, EDMONTON, AB,
T6G 2W2, CANADA

Meddaugh, William, ChevronTexaco Energy Tech. Co., 4800 Fournace Place, P O Box
430, BELLAIRE, TX, 77401, USA

Merchan, Sergio, Encana, 421-7th Avenue SW, Calgary, AB, T2P 4K9, CANADA

Monestiez, Pascal, INRA, Domaine St Paul Site Agropic, Avignon, AVIGNON, 84914,
FRANCE

Murphy, Mark, Snowden Mining Ind. Consultant, P O Box 77, West Perth,, 87 Colin St, WA, 6872, PERTH, 6005, AUSTRALIA

Myers, Donald, University of Arizona, Dept. Of Mathematics, 617 North Santa Rita, TUCSON, AZ, 85721, USA

Naveau, Philippe, University of Colorado, Applied Mathematics Dept., 526 U C B, BOULDER, CO, 80309, USA

Nel, Stefanus, De Beers Consolidated Mines, Private Bag 1, Southdale, Gauteng, ZA 2135, SOUTH AFRICA

Neufeld, Chad, University of Alberta, Department. Of Civil & Environment Engineering, 3-133 Markin/CNRL Natural Resources Engineering Facility, EDMONTON, AB, T6G 2W2, CANADA

Nicholas, Grant D., De Beers, The Dtc, Mendip Court, Bath Rd, South Horrington, WELLS, BA5 3DG, UK

Norrena, Karl, Nexen Canada Ltd, 801-7th Ave SW, CALGARY, AB, T2P 3P7, CANADA

Norris, Brett, Paramount Energy Trust, 500, 630-4th Ave SW, CALGARY, AB, T2P 0L9, CANADA

Nowak, Marek, Nowak Consultants Inc., 1307 Brunette Ave, COQUITLAM, BC, V3K 1G6, CANADA

Okabe, Hiroshi, Imperial College London, Japan Oil, Gas & Metals Nat.co, 1-2-2 Hamada, Mihama-ku, CHIBA- SHI, 261-0025, JAPAN

Ortiz Cabrera, Julian, Universidad de Chile, Av. Tupper 2069, SANTIAGO, 837-0451, CHILE

Osburn, William, St John River Water Management, 4049 Reid St., P O Box 1429, PALATKA, FL, 32178, USA

Parker, Harry, AMEC Inc, 19083 Santa Maria Ave, CASTRO VALLEY, CA, 94546, USA

Perron, Gervais, Mira Geoscience Ltd, 310 Victoria Avenue, Suite 309, WESTMOUNT, QC, H3Z 2M9, CANADA

Pilger, Gustavo, Federal Univ of Rio Grande do, Av. Osvaldo Aranha 99/504, PORTO ALEGRE, 90035-190, BRAZIL

Pintore, Alexandre, Imperial College of London, 11 Walton Well Road, OXFORD, OX2 6ED, UK

Pontiggia, Marco, ENI E&P, Via Emilia 1-5 Pal., Uff.-5013 E, SAN DONATO MILANESE, 20097, ITALY

Porjesz, Robert, CGG, Casa Bote B, Casa 332, LECHERIA, VENEZULA

Potts, Amanda, University of Alberta, Department of Civil & Environmental Engineering, 3-133 Markin/CNRL Natural Resources Engineering Facility, EDMONTON, AB, T6G 2W2, CANADA

Prins, Chris, The DTC, Mendip Court, Bath Road, Minrad, Wells, SOMERSET, BA5 3DG, UK

Pyrcz, Michael, University of Alberta, Dept. Of Civil & Environmental Engineering, 3-133 Markin/CNRL Natural Resources Engineering Facility, EDMONTON, AB, T6G 2W2, CANADA

Remy, Nicolas, Standford University, G E S Department, Braun Hall, STANFORD, CA, 94305, USA

Ren, Weishan, University of Alberta, Dept. Of Civil & Environmental Engineering, 3-133 Markin/CNRL Natural Resources Engineering Facility, EDMONTON, AB, T6G 2W2, CANADA

Riddell, Marla, EnCana, #21, 18-20 Hillcrest Road, LONDON, W5 1HJ, UK

Rivoirard, Jacques, Ecole des Mines de Paris, Centre De Gostatistique, 35 Rue Saint-Honore, FONTAINEBLEAU, 77210, FRANCE

Russo, Ana, CMRP/ IST, Av. Rovisco Pais, 1, Lisboa, 1049-001. PORTUGAL

Saito, Hirotaka, Sandia National Lab., P O Box 5800 M S 0735, ALBUQUERQUE, NM, 87185, USA

Saldanha, Paulo, 4 Dom Thomas Murphy St., WY, BRAZIL

Samal, Abani, Southern Illinois University, 2000 Evergreen Terrace Dr. W., Apt 07, CARBONDALE, IL, 62901, USA

Savoie, Luc, Candian Natural Resources Ltd, 2500, 855-2nd St SW, CALGARY, AB, T2P 4J8, CANADA

Schirmer, Patrice, TOTAL, 2 Place Coupole, La Defense, PARIS, .092604, FRANCE

Schnetzler, Emmanuel, Statios, 1345 Rhode Island Street, SAN FRANCISCO, CA, 94107, USA

Schofield, Neil, Hellman & Schofield Pty. Ltd, P. O. Box 599, Beecroft, Nsw, 2119 , Suite 6, 3 Trelawney St, EASTWOOD, 2122, AUSTRALIA

Scott, Anthony, Placer Dome, P O Box 49330 Bentall Station, Suite 1600-1055
Dunsmuir St., VANCOUVER, BC, V6G 3J3, CANADA

Seibel, Gordon, AngloGold, P O Box 191, VICTOR, CO, 80860, USA

Shi, Genbao, Landmark Graphics, Two Barton Skyway, 1601 S. Mopac Expressway,
AUSTIN, TX, 78746, USA

Skorstad, Arnem Norwegian Computing Center, P O Box 114 Blindern, Gaustadalleen
23, OSLO, NO0314, NORWAY

Soares, Amilcar, CMRP/ IST, Av. Rovisco Pais, 1, LISBOA, 1049-001, PORTUGAL

Soulie, Michel, Ecole Polytechnique, P. O. Box 6079, Station Centre-ville,
MONTREAL, QC, H3C 3A7, CANADA

Srinivasan, Sanjay, University of Texas at Austin, Petroleum & Geosystems Eng., 1
University Station C0300, AUSTIN, TX, 78712, USA

Srivastava, Mohan, FSS Canada, 42 Morton Road, TORONTO, ON, M4C 4N8,
CANADA

Stavropoulos, Achilles, Canadian Natural Resources Ltd, 900, 311-6th Ave. SW,
CALGARY, AB, T2P 3H2, CANADA

Stephenson, John, Imperial College, Dept. Of Earth Science & Eng, Royal School Of
Mines, LONDON, SW7 2AZ, UK

Strebelle, Sebastien, ChevronTexaco ETC, 6001 Bollinger Canyon Rd,, Room D1200,
SAN RAMON, CA, 94583, USA

Suzuki, Satomi, Stanford University, 367 Panama St., STANFORD, CA, 94305, USA

Syversveen, Anne Randi, Norwegian Computing Center, P O Box 411 Blindern, OSLO,
0314, NORWAY

Thurston, Malcolm, De Beers Canada, 65 Overlea Blvd, Suite 400, TORONTO, ON,
M4H 1P1, CANADA

Tjelmeland, Haakon, Norwegian University, Of Science & Technology, Dept. Of
Mathematical Sciences, TRONDHEIM, 7491, NORWAY

Tolmay, Leon, Gold Fields, 4 Cedar Avenue, WESTONARIA, 1779, SOUTH AFRICA

Toscano, Claudio, ENI S.P.A -E & P Division, Via Emilia 1, San Donato Milanese,
MILAN, 1-20097, ITALY

Tran, Thomas, Chevron Texaco, 1546 China Grade Loop, Room A7, BAKERSFIELD, CA, 93308, USA

Tureyen, Omer Inanc, Stanford University, 367 Panama St., 065 Green Earth Sciences Bldg., STANFORD, CA, 94305, USA

Vann, John, Quantitative Geoscience Pty Lt, P O Box 1304, FREMANTLE, 6959, AUSTRALIA

Vasquez, Christina, XU Power, 3010 State St. #309, DALLAS, TX, 75204, USA

Verly, Georges, Placer Dome, P. O. Box 49330, Bentall Station, VANCOUVER, BC, V7X 1P1, CANADA

Voelker, Joe, Stanford University, P O Box 12864, STANFORD, CA, 94309, USA

Wackernagel, Hans, Ecole des Mines De Paris, 35 Rue Saint Honore, Centre De Geostatistique, FONTAINEBLEAU, F-77305, FRANCE

Wagner, Jayne, De Beers, Private Bag X01, Southdale, 2135 Cornerstone Bldg., JOHANNESBURG, 2135, SOUTH AFRICA

Wain, Anthony, Talisman Energy Inc., 888-3rd Street SW, Suite 3400, CALGARY, AB, T2P 5C5, CANADA

Walls, Elizabeth, Petro-Canada, 208, 930 18th Ave SW, CALGARY, AB, T2T 0H1, CANADA

Watson, Michael, Husky Energy, Box 6525, Station D, 707 8th Ave. S W, CALGARY, AB, T2P 3G7, CANADA

Wawruch, Tomasz, Anglo American Chile, Mailbox 16178, Correo 9, 291 Pedro De Valdivia Ave, SANTIAGO, 6640594, CHILE

Wen, Xian- Huan, Chevron Texaco ETC, 6001 Bollinger Canyon Road, D2092, SAN RAMON, CA, 94583, USA

Wu, Jianbing, SCRF, Stanford University, Petroleum Engineering Dept, 367 Panama Street, STANFORD, CA, 94305, USA

Yao, Tingting, ExxonMobil., Upstream Research Company, P. O. Box 2189, S W 508, HOUSTON, TX, 77252, USA

Yarus, Jeffrey, Quantitative Geosciences, LLP, 2900 Wilcrest Suite 370, HOUSTON, TX, 77042, USA

Zanon, Stefan, University of Alberta, Dept. Of Civil & Environmental Engineering,
3-133 Markin/CNRL Natural Resources Engineering Facility, EDMONTON, AB,
T6G 2W2, CANADA

Zhang, Linan, University of Alberta, Dept. Of Civil & Environmental Engineering, 3-133
Markin/CNRL Natural Resources Engineering Facility, EDMONTON, AB, T6G 2W2,
CANADA

Zhang, Tuanfeng, Stanford University, 113 A, E. V, STANFORD, CA, 94305, USA

PLENARY

ACCOUNTING FOR GEOLOGICAL BOUNDARIES IN GEOSTATISTICAL MODELING OF MULTIPLE ROCK TYPES

PAULA LARRONDO and CLAYTON V. DEUTSCH
Department of Civil & Environmental Engineering, 220 CEB,
University of Alberta, Canada, T6G 2G7

Abstract. Geostatistical simulation makes strong assumptions of stationarity in the mean and the variance over the domain of interest. Unfortunately, geological nature usually does not reflect this assumption and we are forced to subdivide our model area into stationary regions that have some common geological controls and similar statistical properties. This paper addresses the significant complexity introduced by boundaries. Boundaries are often soft, that is, samples near boundaries influence multiple rock types.

We propose a new technique that accounts for stationary variables within rock types and additional non-stationary factors near boundaries. The technique involves the following distinct phases: (i) identification of the rock types and boundary zones based on geological modeling and the timing of different geological events, (ii) optimization for the stationary statistical parameters of each rock type and the non-stationary mean, variance and covariance in the boundary zones, and (iii) estimation and simulation using non-stationary cokriging. The resulting technique can be thought of as non-stationary cokriging in presence of geological boundaries.

The theoretical framework and notation for this new technique is developed. Implementation details are discussed and resolved with a number of synthetic examples. A real case study demonstrates the utility of the technique for practical application.

1 Introduction

The most common geostatistical techniques, such as kriging and Gaussian/indicator simulation, are based on strong assumptions of stationarity of the estimation domains. In particular, they are based in a second order stationary hypothesis, that is, the mean, variance and covariance remain constant across the entire domain and they do not depend on the location of the support points but only in the distance between them.

Once estimation domains have been selected, the nature of the boundaries between them must be established. Domain boundaries are often referred to as either 'hard' or 'soft'. Hard boundaries are found when an abrupt change in the mean or variance occurs at the contact between two domains. Hard boundaries do not permit the interpolation or extrapolation across domains. Contacts where the variable changes transitionally across

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 3-12.

the boundary are referred as soft boundaries. Soft domain boundaries allow selected data from either side of a boundary to be used in the estimation of each domain.

It is rather common that soft boundaries are characterized by a non-stationary behavior of the variable of interest in the proximities of the boundary, that is, the mean, variance or covariance are no longer constant within a zone of influence of one rock type into the other, and their values depends on the location relative to the boundary. An example is the increased frequency of fractures towards a boundary between geological domains of structural nature. Faults or brittle zones are examples of this transition. The fractures may cause the average to increase close to the boundary. The increase in the presence of fractures will often lead to an increase in the variance closer to the boundary.

Although soft boundaries are found in several types of geological settings due to the transitional nature of the geological mechanisms, conventional estimation usually treats the boundaries between geological units as hard boundaries. This is primarily due to the limitations of current estimation and simulation procedures. We will show that non-stationary features in the vicinity of a boundary can be parameterized into a local model of coregionalization. With a legitimate spatial model, estimation of grades can be performed using a form of non-stationary cokriging. This proposal provides an appealing alternative when complex contacts between different rock types exist. We develop the methodology in the context of mining geostatistics, but it is widely applicable in many different settings.

2 Theoretical Background

The technique involves the identification of stationary variables within each rock type and additional non-stationary components near boundaries for the mean, variance and covariance. For a geological model with K rock types or estimation domains, there are a maximum of $K(K-1)/2$ boundary zones to be defined. Then, the continuous random function $\mathrm{Z}(\mathbf{u})$ that represents the distribution of the property of interest can be decomposed into K stationary random variables $\mathrm{Z}_{\mathrm{k}}(\mathbf{u})$ $k=1,\ldots,K$ and a maximum of $K(K-1)/2$ non-stationary boundary variables $\mathrm{Z}_{\mathrm{kp}}(\mathbf{u})$, with $k,p=1,\ldots K$ and $\mathrm{Z}_{\mathrm{kp}}(\mathbf{u})=\mathrm{Z}_{\mathrm{pk}}(\mathbf{u})$ (Figure 1). By definition, the non-stationary variable will take values only for locations within the maximum distance of influence of rock type k into rock type p.

The maximum distance of influence orthogonal to the boundary of rock type k into rock type p is denoted $dmax_{kp}$. A boundary zone is defined by two distances: $dmax_{kp}$ and $dmax_{pk}$, since there is no requirement that the regions on each side of the boundary are symmetric, that is, $dmax_{kp}\neq dmax_{pk}$. The modeler using all geological information available and his expertise should establish these distances.

When more than two rock types converge at a boundary, two or more rock types may influence the boundary zone in the adjacent domain. In this case, precedence or ordering rules should determine the dominant boundary zone. Although the behavior of a property near a boundary could be explained by the overlapping of different geological controls, the task of identifying the individuals effects of each rock type and their interactions can be quite difficult. Geological properties are not usually additive and therefore the response of a combination of different rock types is complex. Only one non-stationary factor will be considered at each location. The modeler should put together the precedence rules based on the geology of the deposit. The relative timing of intrusion, deposition or mineralisation events, geochemistry response of the protolith to an alteration or mineralisation process could be used to resolve timing and important variables. If the geological data does not provide sufficient information to establish a geological order of events, a neutral arrangement can be chosen. In this case, the precedent rock type *p* at a location will be the one to which the distance to the boundary is the minimum over all surrounding rock types.

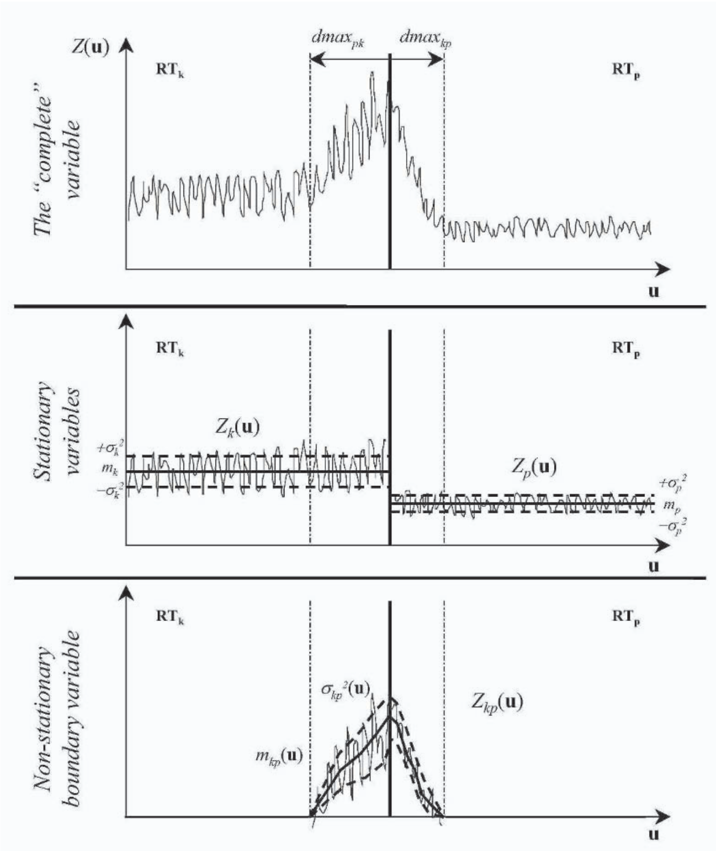


Figure 1: Decomposition of a one-dimensional random function Z(u) in two stationary variables Z_k(u) and Z_p(u), with constant mean and variance, and a non-stationary boundary variable Z_{kp}(u), with a mean and variance that are functions of the distance to the boundary.

STATIONARY AND NON-STATIONARY STATISTICAL PARAMETERS

The mean function of the continuous random function Z(**u**) for a specific rock type *k* will be the mean of the stationary variable Z_k(**u**) plus the mean of any corresponding non-stationary variable Z_{kp}(**u**). The stationary component of the mean (m_k) is independent of location and is a constant value. The non-stationary component of the mean (m_{kp}) is a function of the distance to the boundary, d_{pk}(**u**) and takes values different than zero for locations within the boundary zone defined by rock types *k* and *p*. The mean of rock type *k* is:

$$E\left\{Z(\mathbf{u}_i)\right\}=\begin{cases} m_k & \text{, if } d_{pk}(\mathbf{u}_i)\geq dmax_{pk} \\ m_k+f(d_{pk}(\mathbf{u}_i)) & \text{, otherwise}\end{cases} \qquad \text{where } \mathbf{u}_i\in \mathbf{RT_k}$$

where *p* is the adjacent rock type that shares a boundary with rock type *k* and $f(\bullet)$ is an arbitrary function that describes the mean as a function of distance to the boundary.

Similarly, the variance of Z(**u**) for rock type *k* will be the sum of a constant stationary variance (σ_k^2) due to Z_k(**u**) and the independent non-stationary variance (σ_{kp}^2) due to Z_{kp}(**u**). The variance of a random function Z(**u**) in a rock type *k* is:

$$E\left\{\left(Z(\mathbf{u}_i)-E\left\{Z(\mathbf{u}_i)\right\}\right)^2\right\}=\begin{cases} {\sigma_k}^2 & \text{, if } d_{pk}(\mathbf{u}_i)\geq dmax_{pk} \\ {\sigma_k}^2+g(d_{pk}(\mathbf{u}_i)) & \text{, otherwise}\end{cases} \qquad \text{where } \mathbf{u}_i\in \mathbf{RT_k}$$

where *p* is the adjacent rock type that shares a boundary with rock type *k* and $g(\bullet)$ is an arbitrary function that describes the variance as a function of distance to the boundary.

As with the mean and variance, the covariance structure between two rock types that share a local non-stationary boundary consists of a stationary and a non-stationary component.

$$Cov_Z(\mathbf{u}_i,\mathbf{v}_i)=E\left\{\left(Z(\mathbf{u}_i)-m(\mathbf{u}_i)\right)\cdot\left(Z(\mathbf{v}_i)-m(\mathbf{v}_i)\right)\right\}=Cov_Z^{\mathbf{S}}(\mathbf{h})+Cov_Z^{\mathbf{NS}}(\mathbf{u}_i,\mathbf{v}_i)$$

where **h**=$\mathbf{u}_i$ – $\mathbf{v}_i$. Since Z_k(**u**) and Z_{kp}(**u**) are independent random variables, the cross terms are zero, therefore the covariance of Z(**u**) is the sum of the stationary and non-stationary components. The combination of these components corresponds to a local linear model of coregionalization.

The stationary component of the covariance can be calculated and modeled from data pairs within the internal stationary portion of a rock type, that is $\mathbf{u}_i$ and $\mathbf{v}_i$ belong to rock type *k*, and do not belong to any boundary zone.

To obtain the non-stationary component of the covariance model we will assume that the shape of the spatial correlation of the non-stationary variable Z_{kp}(**u**) *k,p=1,…,K* is stationary and that it can be specified by the modeler. Due to the non-stationary nature of variable Z(**u**) at the boundary zone, this relative stationary spatial model has to by scaled at each point by a non-stationary mean and variance. The relative standardized variogram model for the boundary zone is:

$$\hat{\gamma}_{kp}(\mathbf{u}_i,\mathbf{v}_i)=\frac{1}{2}\cdot E\left\{\left[\frac{Z(\mathbf{u}_i)-m(\mathbf{u}_i)}{\sigma(\mathbf{u}_i)}-\frac{Z(\mathbf{v}_i)-m(\mathbf{v}_i)}{\sigma(\mathbf{v}_i)}\right]^2\right\}$$

where $m(\mathbf{u})=m_{kp}(\mathbf{u})+m_k$ and $\sigma(\mathbf{u})=\sigma_{kp}(\mathbf{u})+\sigma_k$. Expanding and reordering the terms of the squared difference, and since $E\left\{Z(\mathbf{u}_i)^2\right\}=\sigma(\mathbf{u}_i)^2+m(\mathbf{u}_i)$ and $E\left\{Z(\mathbf{u}_i)\right\}=m(\mathbf{u}_i)$, the previous expression becomes:

$$\hat{\gamma}_{kp}(\mathbf{u}_i,\mathbf{v}_i)=1-\frac{Cov_z^{\mathbf{NS}}(\mathbf{u}_i,\mathbf{v}_i)}{\sigma(\mathbf{u}_i)\cdot\sigma(\mathbf{v}_i)}$$

Reordering the terms and replacing the mean and variance by the sum of their stationary and non-stationary components, we obtain an expression for the non-stationary covariance model:

$$\begin{aligned}Cov_Z^{\mathbf{NS}}(\mathbf{u}_i,\mathbf{v}_i)&=E\left\{Z(\mathbf{u}_i)\cdot Z(\mathbf{v}_i)\right\}-(m_{kp}(\mathbf{v}_i)+m_k)\cdot(m_{kp}(\mathbf{u}_i)+m_k)\\&=(1-\hat{\gamma}_{kp}(\mathbf{u}_i,\mathbf{v}_i))\cdot(\sigma_{kp}(\mathbf{u}_i)+\sigma_k)\cdot(\sigma_{kp}(\mathbf{v}_i)+\sigma_k)\end{aligned}$$

Currently we assume that the shape, anisotropies and nugget effect of the relative standardized variogram are inputs from the modeler; only the range must be established through an optimization algorithm.

3 Optimization of the Statistical Parameters

We need to find the optimum $f(d_{pk}(\mathbf{u}_i))$, $g(d_{pk}(\mathbf{u}_i))$ and $Cov_Z^{\mathbf{NS}}(\mathbf{u}_i,\mathbf{v}_i)$ that fit the distribution of the random variable Z(**u**) at the boundary zone given the stationary components of mean, variance and covariance, a set of precedence rules and the maximum distances of influence within the rock type model.

We will consider that the non-stationary components of the mean and variance follow a linear function of the distance to the boundary (d_{pk}). In this scenario, the optimization of the parameter m_{kp} and σ_{kp}^2 will be equivalent to optimizing estimates of the intercepts at zero distance to boundary: a_{kp} and b_{kp}, considering $a_{kp}= a_{pk}$ and $b_{kp}= b_{pk}$.

The mean m_{kp} is optimized given that m_k is known from the experimental average of data within rock type k, outside any boundary zone. The objective function is:

$$O_m = \sum_{k=1}^{K}\sum_{p=1}^{P}\sum_{i=1}^{N_{kp}}\left[z(\mathbf{u}_i) - (\hat{m}_k + m_{kp}(\mathbf{u}_i))\right]^2$$

where z(**u***ᵢ*) is the outcome value at every data location in the boundary zone, N_{kp} is the total number of data in zone *k-p*, $\hat{m}_k$ is the experimental average of all data in $\mathbf{RT_k}$ and outside any boundary zone, and $m_{kp}(\mathbf{u}_i)$ is the non-stationary mean at location $\mathbf{u}_i$ calculated as:

$$m_{kp}(\mathbf{u}_i) = \begin{cases} \dfrac{\left(dmax_{kp} - d_{kp}(\mathbf{u}_i)\right)}{dmax_{kp}} \cdot a_{kp} & \text{for } 0 \le d_{kp}(\mathbf{u}_i) \le dmax_{kp} \\ \dfrac{\left(dmax_{pk} - d_{pk}(\mathbf{u}_i)\right)}{dmax_{pk}} \cdot a_{kp} & \text{for } 0 \le d_{pk}(\mathbf{u}_i) \le dmax_{pk} \\ 0 & \text{for } d_{kp}(\mathbf{u}_i) \ge dmax_{kp} \text{ and } d_{pk}(\mathbf{u}_i) \ge dmax_{pk} \end{cases} \qquad (1)$$

The optimization of the mean can be achieved by iteratively modified a_{kp} $\forall k,p$, in a random fashion while accepting all changes in a_{kp} that reduce the objective function. This is a simplified version of the simulated annealing formalism.

The optimum $\sigma_{kp}{}^2$, will be the one that minimizes the following objective function:

$$O_{\sigma^2} = \sum_{k=1}^{K}\sum_{p=1}^{P}\sum_{i=1}^{N_{kp}}\left[\mathrm{r}(\mathbf{u}_i)^2 - (\hat{\sigma}_k{}^2 + \sigma_{kp}{}^2(\mathbf{u}_i))\right]^2$$

where r($\mathbf{u}_i$) is the residual value at every location in the boundary zone. $\hat{\sigma}_k{}^2$ is the experimental variance of all data within the stationary region of rock type *k* and $\sigma_{kp}{}^2(\mathbf{u}_i)$ is the non-stationary variance at location $\mathbf{u}_i$ calculated from a linear expression for the intercept b_{kp} similar to Equation 1.

Figure 2 shows the stationary and non-stationary mean and variance for a 1D synthetic example. The optimum intercepts a_{kp} and b_{kp} are in agreement with the reference.

To find the optimum covariance model we minimize the following objective function:

$$O_{Cov} = \sum_{i=1}^{N}\left[\hat{C}\left(z(\mathbf{u}_i), z(\mathbf{v}_i)\right) - C_{MOD}\left(z(\mathbf{u}_i), z(\mathbf{v}_i)\right)\right]^2$$

where $\hat{C}$ denote the experimental covariance of the pair located at $\mathbf{u}_i$ and $\mathbf{v}_i$, which is just the multiplication of the two residual values: $r(\mathbf{u}_i)\cdot r(\mathbf{v}_i)$, and C_{MOD} the modeled boundary covariance, corresponding to the sum of the stationary and non-stationary component.

Finding the optimum covariance model of a boundary zone is equivalent to optimizing the range of the relative standardized variogram scaled by the non-stationary standard deviation. The range is iteratively modified by a random amount until the difference between the experimental and modeled covariance is minimized. For this 1D example, the optimum range of the non-stationary covariance structure (Figure 3) is 6.4 meters, acceptably similar to the 10 meters range of the variogram used to obtain the reference.

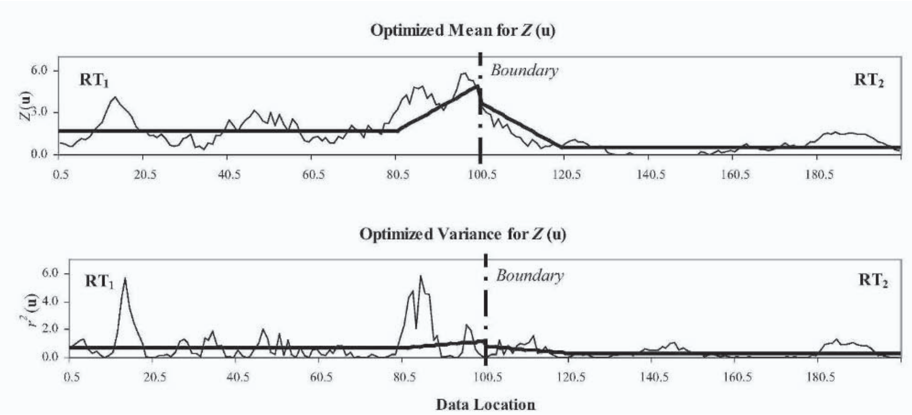


Figure 2: 1D example stationary and optimized non-stationary mean and variance.

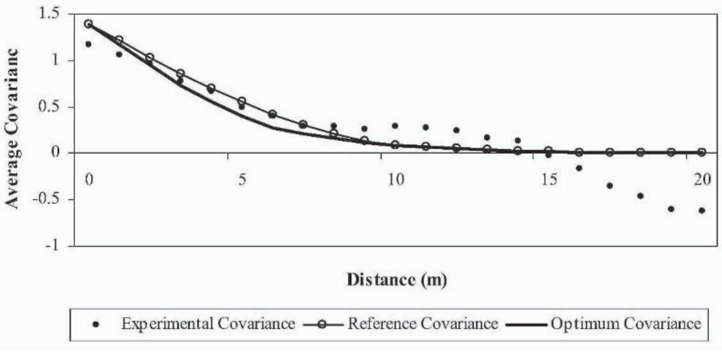


Figure 3: Optimum non-stationary covariance of 1D example (solid line), experimental covariance from pairs within the boundary zone (dots) and original covariance of the non-stationary component used to build the synthetic dataset (line/dots).

4 Estimation in presence of local non-stationary boundaries

The basic linear regression equation for non-stationary simple cokriging is:

$$z^*(\mathbf{u}) - m(\mathbf{u}) = \sum_{\alpha=1}^{n} \lambda_\alpha(\mathbf{u}) \cdot \left[z(\mathbf{u}_\alpha) - m(\mathbf{u}_\alpha) \right]$$

where $z^*(\mathbf{u})$ is the estimate at unsampled location $\mathbf{u}$, $m(\mathbf{u})$ is the stationary plus the non-stationary mean value at location $\mathbf{u}$, $\lambda_\alpha(\mathbf{u})$ is the weight assigned to datum $z(\mathbf{u}_\alpha)$, n is the number of close data to the location $\mathbf{u}$ being estimated, and $m(\mathbf{u}_\alpha)$ are the n stationary plus the non-stationary mean values at the data locations.

To find the optimal weights $\lambda_\alpha(\mathbf{u})$, $\alpha=1,\ldots,n$ the kriging system must be solved:

$$\sum_{\beta=1}^{n} \lambda_\beta(\mathbf{u}) \ \cdot Cov(\mathbf{u}_\alpha, \mathbf{u}_\beta) = Cov(\mathbf{u}, \mathbf{u}_\alpha) \quad \text{with } \alpha, \beta = 1, \ldots, n$$

where $\lambda_\alpha(\mathbf{u})$, $\alpha=1,.., n$ are the simple kriging weights, $Cov(\mathbf{u}_\alpha, \mathbf{u}_\beta)$, $\alpha,\beta=1,.., n$ correspond to the data-to-data covariances, and $Cov(\mathbf{u}, \mathbf{u}_\alpha)$, $\alpha=1,.., n$ are the data-to-unknown location covariances. In the presence of local non-stationary boundaries, the terms of the data covariance matrix and the vector of data-to-estimate covariances are obtained combining the stationary and non-stationary covariance model components. If both locations are in the same rock type and both are in the same boundary zone, the covariance is the stationary plus the non-stationary covariances; otherwise, it is only the stationary component. If they are in different rock types and both samples are in the same boundary zone the covariance is the non-stationary component only. The covariance is zero in all other cases.

For the 1D example, the kriging estimates reproduce well the reference, using as conditioning data one of four grid nodes of the reference (Figure 4).

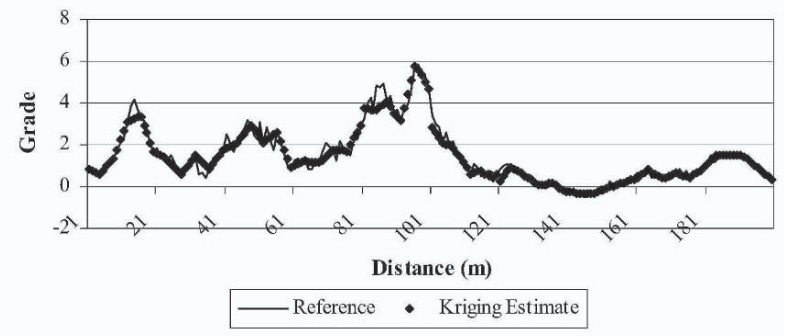


Figure 4: Reference versus kriging estimates, 1D example.

5 Application

The rock type model of a porphyry copper deposit in Northern Chile was used to create a reference image with simulated grades (Figure 5). This reference image was sampled in a 100x100 meters grid. The geological model has five rock types and six non-stationary soft boundaries.

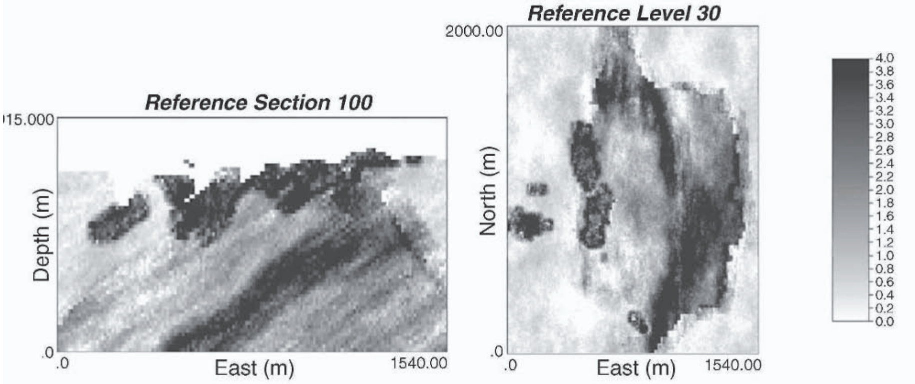


Figure 5: Section and level maps of the reference used for the 3D application.

The reference intercepts for the non-stationary mean and variance are well reproduced by the optimization subroutines for all boundary zones, as well as the optimum ranges compared to the range used in the transformed unconditional simulation.

The correlation between the estimates and the reference value is around 0.8 for each boundary zone. The reference stationary means of each rock type is reproduced almost exactly by the kriging estimation. The variance of the estimates is lower than the reference, which is expected since kriging has a smoothing effect. The non-stationary behavior of the mean is also very well reproduced by the proposed non-stationary kriging as shown in Figure 6. Although the variance of the estimates in the boundary zone is lower than the reference, as expected, the increasing trend toward the boundary is well reproduced.

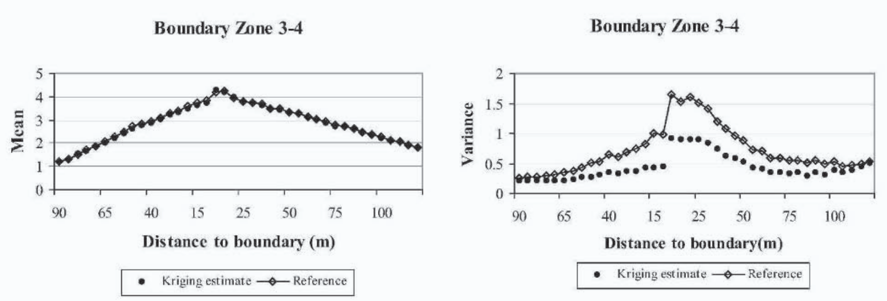


Figure 6: Mean and variance of the kriging estimates versus the reference image at the one of non-stationary boundary zone. Each point corresponds to the average/variance of all grid nodes within a 5 meters interval of the distance to the boundary.

Cross validation results show that the data are reliably estimated both in the stationary and the non-stationary regions. In particular, for all data within the non-stationary regions, if compared with ordinary kriging using a typical soft boundary approach, the proposed methodology shows a higher coefficient of correlation (Figure 7).

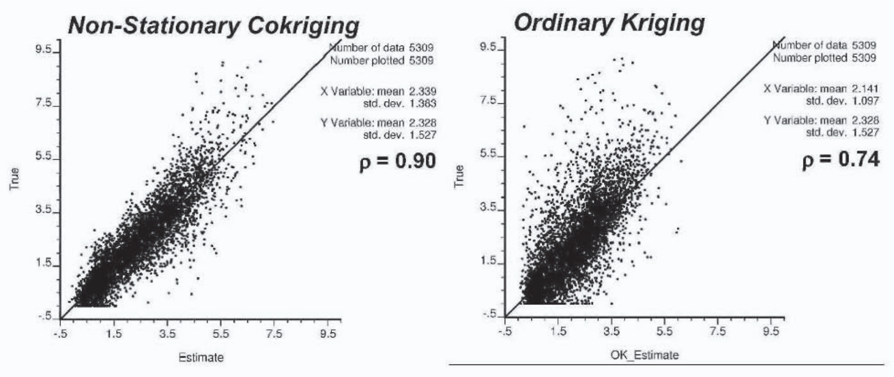


Figure 7: Cross validation comparison between the proposed methodology, non-stationary cokriging, and ordinary kriging with soft boundaries.

6 Conclusions

This new technique provides a theoretically robust methodology to handle non-stationary soft boundaries. The non-stationary features of the mean, variance and covariance are parameterized into a legitimate local model of coregionalization. Through this spatial model a non-stationary form of cokriging accounts for the changes in mean and variance at the vicinity of boundaries. The kriging estimates reproduce the non-stationary behavior of the conditioning data at the geological contacts, and it also reproduces the stationary means of each rock type in the model. A decrease in the global variance is due to the smoothing effect of kriging.

By construction, the kriging variance also has a non-stationary component. Since the kriging variance is the missing variability that is reintroduced in simulation, its implementation in the presence of local non-stationary boundaries will be delicate and is part of the future work.

Acknowledgements

We would like to acknowledge the industry sponsors of the Centre for Computational Geostatistics at the University of Alberta for supporting this research. We also thank Canada’s National Sciences and Engineering Research Council (NSERC) for funding.

DATA INTEGRATION USING THE PROBABILITY PERTURBATION METHOD

JEF CAERS
Department of Petroleum Engineering,
Stanford University, Stanford, CA 94305-2220, USA

Abstract. A new method, termed probability perturbation, is developed for solving non-linear inverse problem under a prior model constraint. The method proposed takes a different route from the traditional Bayesian inverse models that rely on prior and likelihood distribution for stating, then sampling, from a posterior distribution. Instead, the probability perturbation method relies on so-called pre-posterior distributions, which state the distribution of the unknown parameter set given each individual data type (linear or non-linear). Sampling consists of perturbing the probability models used to generate the model realization, by which a chain of realizations is created that converge to match any type of data. The probability perturbations are such that the underlying spatial structure (prior model) of the stochastic algorithm is maintained through all perturbations. A simple example illustrates the approach.

1 Introduction

Conditioning stochastic simulations is of the utmost importance in many applications of geostatistics. Most of the current algorithms can condition to data that are linear or pseudo-linear (i.e. linearized using transformations) and of a single-point nature, by which it is understood that there is a linear relationship between data and the unknown taken one at the time. For example, the technique of sequential simulation, either under Gaussian or non-Gaussian assumptions, can be conditioned to hard data, (pseudo-linear) block average data or soft data, the latter through some form of (linear) co-kriging.

Many applications of geostatistics call for the inclusion of non-linear and multiple-point data. The relationship between data and unknown is provided through a complex multi-dimensional transfer function, also termed a forward model. This function often is modelled numerically through a partial differential equation (or its numerical implementation) such as in aquifer models, pollution models, ecological models and for models of flow in oil & gas reservoirs. Integrating this type of data into stochastic simulation calls for an iterative solution (trial-and-error) of an often ill-posed inverse problem. Sampling solutions such a Markov chain Monte Carlo within a Bayesian framework have been proposed (Mosegaard and Tarantola, 1995; Omre and Tjelmeland, 1997) but are often prohibitive in terms of CPU when the forward model is expensive to compute.

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 13-22.

In this paper a new and practical approach within the context of Bayesian inverse modelling is presented. The method allows to condition stochastic simulations to virtually any type of non-linear data. The principle of this method is simple: by perturbing the probabilities models used to generate the model realization, a chain of realizations is created that converge to match any type of data. It is shown that the probability perturbations are such that the underlying spatial structure of the stochastic algorithm is maintained through all perturbations. The probability perturbations can be parameterized by a single parameter or by multiple parameters in order to provide enough flexibility to match large models with a possible large set of non-linear data.

2 Bayesian inverse modelling

Inverse modeling consists of finding a set of model parameters **m** given some data **d**. In the Earth Sciences the model parameters are often unknown material or rock properties located on a 3D grid, e.g. unknown soil type or unknown petrophysical properties in the subsurface. Most inverse problems are underdetermined, meaning that a joint distribution of model parameters is possible given the data. In this paper, we will divide the data into two sets: (1) $\mathbf{d}_1$, or "easy data" which have a simple linear or pseudo-linear relationship with the model parameters, and (2) $\mathbf{d}_2$ or "difficult data" which exhibit a multi-point, non-linear relationship with **m**. For the data $\mathbf{d}_1$, many fast and robust direct sampling methods exist for sampling the distribution of possible model realizations **m**. To condition to data $\mathbf{d}_2$, iterative sampling is required. The posterior distribution from which these samples are drawn is, in a Bayesian context, decomposed into a likelihood and prior distribution

$$f(\mathbf{m} \mid \mathbf{d}_1, \mathbf{d}_2) = \frac{f(\mathbf{d}_1, \mathbf{d}_2 \mid \mathbf{m}) f(\mathbf{m})}{f(\mathbf{d}_1, \mathbf{d}_2)} \simeq \frac{f(\mathbf{d}_1 \mid \mathbf{m}) f(\mathbf{d}_2 \mid \mathbf{m}) f(\mathbf{m})}{f(\mathbf{d}_1, \mathbf{d}_2)} \qquad (1)$$

where the likelihood $f(\mathbf{d}_1, \mathbf{d}_2 | \mathbf{m})$ is further decomposed into $f(\mathbf{d}_1 \mid \mathbf{m})$ and $f(\mathbf{d}_2 | \mathbf{m})$ under the assumption of conditional independence. This assumption makes inference of the likelihood feasible. The assumption of conditional independence is difficult to verify yet may have considerable consequence to the model definition (model for the posterior distribution).

The prior density $f(\mathbf{m})$ describes the dependency between the model parameters. In a spatial context such dependency refers to the spatial structure of **m**. The likelihood density $f(\mathbf{d}|\mathbf{m})$ models the stochastic relationship between the observed data and each particular model **m** retained. This likelihood would account for model and measurement errors. In the absence of any such errors, the data **d** and model **m** are related through a forward model *g*

$$\mathbf{d} = g(\mathbf{m})$$

Markov chain Monte Carlo methods encompass a set of iterative sampling techniques for drawing samples from this posterior distribution. Popular sampling methods are rejection sampling and the Metropolis sampler (Metropolis et al., 1953; Besag and

Green, 1993; Mosegaard and Tarantola, 1995; Omre and Tjelmeland, 1997). These samplers avoid specification of $f(\mathbf{d})$ and are iterative in nature in order to obtain a single sample $\mathbf{m}^{(\ell)}$ of $f(\mathbf{m}|\mathbf{d})$. Generating multiple (conditioned to **d**) samples $\mathbf{m}^{(\ell)}$, ℓ=1,…,L in this manner quantifies the uncertainty modeled in $f(\mathbf{m}|\mathbf{d})$.

While theoretically sounds, there are some important practical limitations to this approach. First, obtaining iterative samples are CPU demanding and may take many thousand of evaluations to converge. This is impractical when the forward model *g* takes a few hours to compute (e.g. flow simulations, solving elastic wave equations). Secondly, for reason of analytical convenience a Gaussian model is often adopted for either likelihood and/or prior distribution. A Gaussian model limits modelling realistic spatial structures on **m**. Moreover, the assumption of conditional independence in Eq. (1) limits the proper modelling of the full dependence between data $\mathbf{d_1}$ and the data $\mathbf{d_2}$. In this paper we propose a method for dealing with both issues: (1) realistic non-Gaussian prior and (2) alternatives to the conditional independence hypothesis.

3 Methodology

3.1 Sampling the prior

To emphasize non-Gaussianity, the methodology will be developed for binary model parameters, although the method works equally well for multi-category and continuous variables. At each location of a 3D grid an unknown model parameter m_i is modelled through a binary indicator variable

$$I(\mathbf{u}_i) = \begin{cases} 1 & if \quad \text{the "event" occurs at } \mathbf{u}_i \\ 0 & else \end{cases}$$

where “event” could represent any spatially distributed phenomenon. The model parameters are then given by the set of binary indicators

$$\mathbf{m} = \{I(\mathbf{u}_1), I(\mathbf{u}_2), \ldots, I(\mathbf{u}_N)\}$$

with joint (prior) distribution

$$f(\mathbf{m}) = \text{Prob}\{I(\mathbf{u}_1) = i(\mathbf{u}_1), I(\mathbf{u}_2) = i(\mathbf{u}_2), \ldots, I(\mathbf{u}_N) = i(\mathbf{u}_N)\}$$

In this paper we will use sequential simulation methods to sample from either prior or posterior distribution, by relying on the following decomposition of the joint distribution

$$f(\mathbf{m}) = \text{Prob}\{I(\mathbf{u}_1) = 1\} \times \text{Prob}\{I(\mathbf{u}_2) = 1 \,|\, i(\mathbf{u}_1)\} \times \ldots \times \text{Prob}\{I(\mathbf{u}_N) = 1 \,|\, i(\mathbf{u}_1), \ldots, i(\mathbf{u}_{N-1})\}$$

Sequential sampling from each of these conditional distribution amounts to sampling from a joint prior distribution. In actual field cases, prior information on **m** comes in the form of limited statistics (e.g. a spatial covariance). The type of multi-variate density *f* always needs to be assumed. In all sequential simulation approaches, except Gaussian simulation, the decision of distribution type is not made on the joint distribution, but on the conditional distributions. An example of such approach is direct sequential simulation (*dssim*, Journel, 1993), where the conditional distribution can be of any type, as long as they have mean and variance provided by a simple kriging system. Another example is *snesim* where the conditional distributions are derived from training images (Strebelle, 2002).

3.2 Sampling the posterior

To sample from the posterior, a similar sequential decomposition approach is considered. For simplicity, the data $\mathbf{d}_1$ constitute direct observations (hard data) of the model parameters at a set of *n* spatial locations, but in general could constitute any linear data,

$$\mathbf{d}_1 = \{i(\mathbf{u}_\alpha), \alpha = 1,\ldots,n\}$$

The relationship between the non-linear data and model parameters is modelled through a forward model *g*

$$\mathbf{d}_2 = g(\mathbf{m}) = g(I(\mathbf{u}_1), I(\mathbf{u}_2),\ldots, I(\mathbf{u}_N))$$

The goal is to draw samples from the joint (posterior) distribution of the model parameters given the two data sets

$$f(\mathbf{m} \mid \mathbf{d}_1, \mathbf{d}_2) = \text{Prob}\{I(\mathbf{u}_1) = i(\mathbf{u}_1), I(\mathbf{u}_2) = i(\mathbf{u}_2),\ldots, I(\mathbf{u}_N) = i(\mathbf{u}_N) \mid \{i(\mathbf{u}_\alpha), \alpha = 1,\ldots,n\}, \mathbf{d}_2\}$$

To make this practically feasible, the following decomposition is used:

$$\begin{aligned} f(\mathbf{m} \mid \mathbf{d}_1, \mathbf{d}_2) = \text{Prob}\{&I(\mathbf{u}_1) = 1 \mid \{i(\mathbf{u}_\alpha), \alpha = 1,\ldots,n\}, \mathbf{d}_2\} \times \\ &\text{Prob}\{I(\mathbf{u}_2) = 1 \mid i(\mathbf{u}_1), \{i(\mathbf{u}_\alpha), \alpha = 1,\ldots,n\}, \mathbf{d}_2\} \times \ldots \times \\ &\text{Prob}\{I(\mathbf{u}_N) = 1 \mid i(\mathbf{u}_1),\ldots, i(\mathbf{u}_{N-1}), \{i(\mathbf{u}_\alpha), \alpha = 1,\ldots,n\}, \mathbf{d}_2\} \end{aligned} \quad (2)$$

Generating a sample of a (not explicitly stated) posterior distribution is equivalent to generating sequential samples from conditional distributions of the type

$$\begin{aligned} &\text{Prob}\{I(\mathbf{u}_j) = 1 \mid i(\mathbf{u}_1),\ldots, i(\mathbf{u}_{j-1}), \{i(\mathbf{u}_\alpha), \alpha = 1,\ldots,n\}, \mathbf{d}_2\} = \text{Prob}(A_j \mid \mathbf{B}_j, \mathbf{C}) \\ &\quad \text{with } A_j = \{I(\mathbf{u}_j) = 1\};\ \mathbf{B}_j = \{i(\mathbf{u}_1),\ldots, i(\mathbf{u}_{j-1}), \{i(\mathbf{u}_\alpha), \alpha = 1,\ldots,n\}\};\ \mathbf{C} = \mathbf{d}_2 \end{aligned} \quad (3)$$

A simpler notation in terms of ‘A’ (unknown), ‘B’ (easy data) and ‘C’ (difficult data) has been used to make further development clear. To further specify the conditionals in Eq. (3), we propose a decomposition of Prob(A_j|$\mathbf{B}_j$,$\mathbf{C}$) into two pre-posteriors Prob(A_j|$\mathbf{B}_j$) and Prob(A_j|$\mathbf{C}$) using Journel’s decomposition (or tau-model, Journel, 2002) of the type

$$\text{Prob}(A_j \mid \mathbf{B}_j, \mathbf{C}) = \frac{1}{1+x} \text{ with } x = b\left(\frac{c}{a}\right)^{\tau} \text{ where:}$$
$$b = \frac{1 - \text{Prob}(A_j \mid \mathbf{B}_j)}{\text{Prob}(A_j \mid \mathbf{B}_j)},\ c = \frac{1 - \text{Prob}(A_j \mid \mathbf{C})}{\text{Prob}(A_j \mid \mathbf{C})},\ a = \frac{1 - \text{Prob}(A_j)}{\text{Prob}(A_j)} \tag{4}$$

Working with pre-posteriors will lead to an approach that is different from a classical Bayesian inversion which would involve the likelihoods Prob($\mathbf{B}_j|A_j$) and Prob($\mathbf{C}|A_j$). This difference will lead to a fundamentally different sampling method as well. Stating "pre-posteriors", instead of likelihoods, allows using (non-iterative) sequential simulation, instead of (iterative) McMC.

The τ-value in Eq. (4) allows modeling explicitly the full dependency between the **B**-data and **C**-data. The case when τ=1 is equivalent to an assumption of standardized conditional independence. In the context of sequential simulation, the pre-posterior Prob($A_j|\mathbf{B}_j$) is simply the conditional distribution of the unknown A_j given any previously simulated nodes. The remaining pre-posterior Prob($A_j|\mathbf{C}$) cannot be directly estimated, instead, a new sampling technique termed probability perturbation is introduced.

3.3 PROBABILITY PERTURBATION

Using sequential simulation, a sample realization can be drawn from the prior model, conditioned to the data $\mathbf{d}_1$. If the pre-posterior Prob($A_j|\mathbf{C}$) were known, then including the data $\mathbf{d}_2$ could be achieved through Eq. (4) and a sequential simulation from the conditionals through Eq (2). Since this is not the case, the initial sample conditioned to $\mathbf{d}_1$, will be used as an initial guess for further matching the data $\mathbf{d}_2$ iteratively. To achieve this, the unknown pre-posterior Prob($A_j|\mathbf{C}$) is modelled using a single parameter model in the following equation:

$$\text{Prob}(A_j \mid \mathbf{C}) = \text{Prob}(I(\mathbf{u}_j) = 1 \mid \mathbf{C}) = (1 - r_C) \times i_B^{(o)}(\mathbf{u}_j) + r_C \times P(A_j), \quad j = 1, \ldots, N \tag{5}$$

where r_C is a parameter between [0,1], not dependent on $\mathbf{u}_j$. $\{i_B^{(0)}(\mathbf{u}_j), j=1,\ldots,N\}$ is an initial realization conditioned to the $\mathbf{d_1}$ data (B-data) only. Given Eq. (5), Prob($A_j|\mathbf{C}$) can be calculated for a given value of r_C and for a given initial realization constrained to the B-data. Next, the probability Prob($A_j|\mathbf{C}$) is combined with Prob($A_j|\mathbf{B}_j$) to form the conditionals Prob($A_j|\mathbf{B}_j,\mathbf{C}$) by which sequential simulation is possible, Eq. (2), and a new realization $\{i^{(1)}(\mathbf{u}_j), j=1,\ldots,N\}$ is generated. The new realization is dependent on the initial realization and the value of r_C. To get some more insight into the role of the value r_C, consider the examples in Figure 1. Each row shows in its first column an initial realization $\{i_B^{(0)}(\mathbf{u}_j), j=1,\ldots,N\}$ generated with different sequential simulation methods. The next columns contain realizations $\{i^{(1)}(\mathbf{u}_j), j=1,\ldots,N\}$ for various values of r_C and using a random seed *s'* different from the random seed used to generate the initial realization. The important message of Figure 1 is that regardless of the value of r_C the each realization honors the same spatial statistics as the initial model, i.e. the prior distribution is maintained.

In case r_C=0, then Prob(A_j|**C**)= $i_B^{(0)}$(**u**$_j$), hence per Eq (4) Prob(A_j|**B**$_j$,**C**)= $i_B^{(0)}$(**u**$_j$). In other words, the initial realization is re-created. In case r_C=1, then Prob(A_j|**C**)=P(A_j), a simple calculation using Eq. (4) shows that in that case Prob(A_j|**B**$_j$,**C**)= Prob(A_j|**B**$_j$). Since the seed *s*' is different from the seed *s*, the realization $\{i^{(1)}(\mathbf{u}_j),\ j=1,\ldots,N\}$ is equiprobable with the initial realization $\{i_B^{(0)}(\mathbf{u}_j), j=1,\ldots,N\}$. In other words, r_C=1 entails a "maximum perturbation" within the prior model constraints.

A value r_C between (0,1) will therefore generate a perturbation $\{i^{(1)}(\mathbf{u}_j,r_C),\ j=1,\ldots,N\}$ between the initial realization and another equiprobable realization both conditioned to the data $\mathbf{d}_1$ and each honoring the prior model statistics. An optimal value for r_C can be picked by selecting the perturbation for which the mismatch between the forward model simulation and actual data $\mathbf{d}_2$, namely

$$O(r_C) = \left\| g(i^{(1)}(\mathbf{u}_j, r_C)) - \mathbf{d}_2 \right\| \qquad (6)$$

is minimal.

3.4 Probability perturbation algorithm

The probability perturbation of the initial realization is likely to reduce the objective function in Eq. (6), however, minimizing $O(r_C)$ would only achieve a local minimum since the perturbation takes place between just two equiprobable realizations. To further reduce the objective function, the perturbations are iterated in the following algorithm:

- choose random seed
- generate an initial realization $i^{(0)}(\mathbf{u}_j), \forall j$
- change random seed
- Until the data $\mathbf{d}_2$ are matched to some desired level
  - Minimize to get r_C^{opt}

    $$O(r_C) = \left\| g(i^{(1)}(\mathbf{u}_j, r_C)) - \mathbf{d}_2 \right\|$$
  - Change random seed
  - Assign

    $$i^{(0)}(\mathbf{u}_j) \leftarrow i^{(1)}(\mathbf{u}_j, r_C^{opt}), \forall j$$

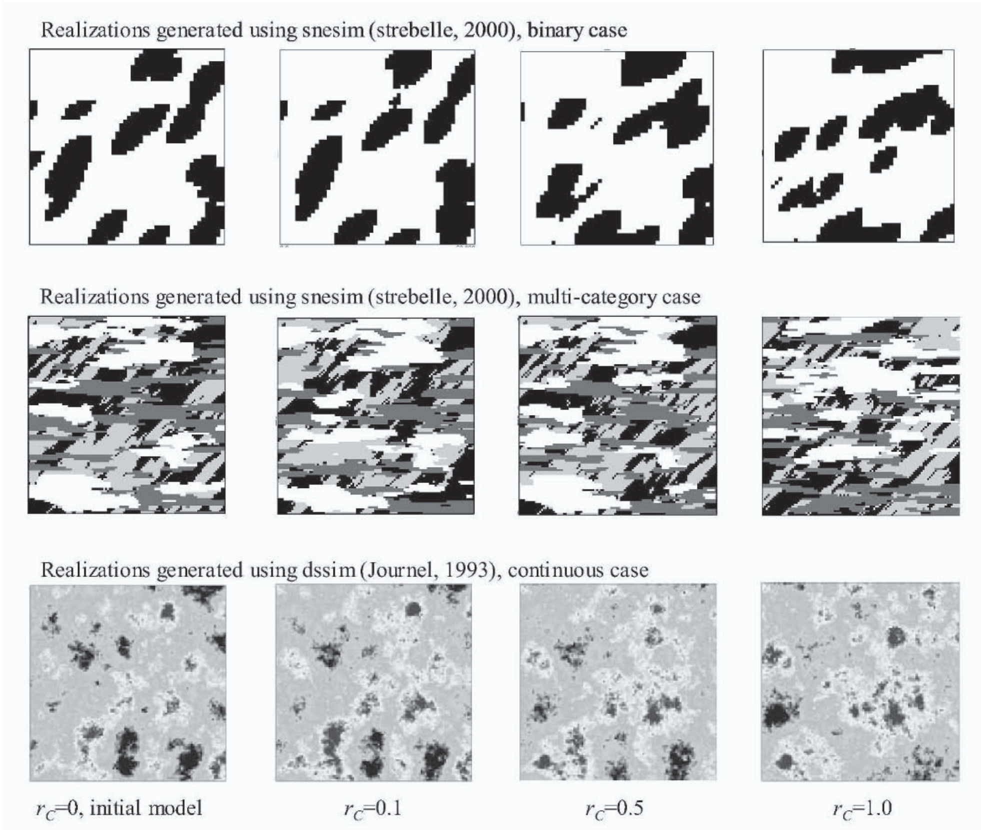

Realizations generated using snesim (strebelle, 2000), binary case
Realizations generated using snesim (strebelle, 2000), multi-category case
Realizations generated using dssim (Journel, 1993), continuous case
r_C=0, initial model
r_C=0.1
r_C=0.5
r_C=1.0

Figure 1: Left picture of each row is an initial guess realization, then followed by perturbation of this initial realization parameterized by a parameter r_C.

3.5 REGIONAL PROBABILITY PERTURBATION

The probability perturbation method generates a perturbation between an initial guess realization and another equiprobable realization that is parameterized using a single parameter. In a spatial context this induces a perturbation of each individual model parameter $i(\mathbf{u}_j)$ that is, *in probability*, the same for all $\mathbf{u}_j$. Parameterizing a perturbation using a single parameter may not effectively solve complex spatial inverse problem.

The above presented method does not restrict a higher order parameterization: the value of r_C can be made dependent on location

$$\text{Prob}(A_j \mid \mathbf{C}) = \text{Prob}(I(\mathbf{u}_j) = 1 \mid \mathbf{C}) = (1 - r_C(\mathbf{u}_j)) \times i_B^{(s)}(\mathbf{u}_j) + r_C(\mathbf{u}_j) \times P(A_j) \quad (7)$$

The use of (7) in the probability perturbation method now requires a multi-dimensional optimization on all $r_C(\mathbf{u}_j)$, j=1,…,N. To avoid a potentially difficult full multi-dimensional search for the best $r_C(\mathbf{u}_j)$, j=1,…,N, a region-wise parameterization of these parameters is proposed. Consider M regions in the domain of study, each region R_m, m=1,…, M consists of a set of grid node locations,

$$R_m = \{\mathbf{u}_i^{(m)}, \mathbf{u}_j^{(m)}, \ldots\}$$

Which nodes belong to which region is a problem specific question. The number of regions M however is likely to be considerably less than the number of grid nodes N. The pre-posterior of Eq. (7) is rewritten using a region-wise parameterization as follows:

$$\mathrm{Prob}(A_j^{(m)} \mid \mathbf{C}) = \mathrm{Prob}(I(\mathbf{u}_j^{(m)}) = 1 \mid \mathbf{C}) = (1 - r_C^{(m)}) \times i_B^{(s)}(\mathbf{u}_j^{(m)}) + r_C^{(m)} P(A_j^{(m)}), \quad j = 1, \ldots, N$$

where the parameter $r_C^{(m)}$ is the same for all grid nodes $\mathbf{u}_j^{(m)}$ of region R_m,. An efficient strategy for jointly optimizing on all M $r_C^{(m)}$ parameters is discussed in Hoffman and Caers (2003).

4 Example

The aim of this paper is to present the inverse theory behind the probability perturbation method which has been extensively researched and applied to real cases in the context of inversion of flow data in oil reservoirs (Caers, 2003; Hoffman and Caers, 2004). We refer the reader to these paper for practical examples.

In this paper, a simple but rather revealing example is presented and illustrated in Figure 2. The model consists of a grid with three nodes, $\mathbf{u}_1$, $\mathbf{u}_2$ and $\mathbf{u}_3$. Each node can be either black, $I(\mathbf{u})$=1 or white, $I(\mathbf{u})$=0. The model $\mathbf{m}$ is therefore simply

$$\mathbf{m} = \{I(\mathbf{u}_1), I(\mathbf{u}_2), I(\mathbf{u}_3)\}$$

The spatial dependency of this simple 1D model is described by a 1D training image shown in Figure 2. One can extract, by scanning the training image with a 3 x 1 template, the prior distribution, $f(\mathbf{m})$, of the model parameters, as shown in Figure 2. To test the probability perturbation method we consider two data: the first datum is a point measurement (B-data, or "easy data") namely, $i(\mathbf{u}_2)$=1 (a black pixel in the middle), the second one is $I(\mathbf{u}_1)$+$I(\mathbf{u}_2)$+$I(\mathbf{u}_3)$=2 (C-data or "difficult data"). The problem posed is:

$$\text{What is } \mathrm{P\,rob}(I(\mathbf{u}_1) = 1 \mid i(\mathbf{u}_2) = 1, I(\mathbf{u}_1) + I(\mathbf{u}_2) + I(\mathbf{u}_3) = 2)\,?$$

or in simple notation P(A|B,C) ?

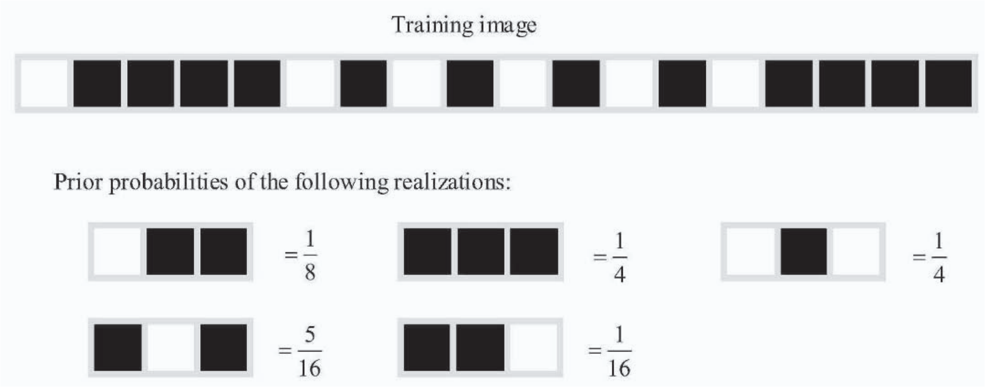


Figure 2: illustrative example: (top) 1D training image (bottom) derived from the training image are the prior probabilities of the model m

To get the answer we use four alternative techniques:

1. Get the true answer by elimination from the prior: $\text{Prob}(A\,|\,B,C)=\dfrac{\frac{1}{16}}{\frac{1}{8}+\frac{1}{16}}=\dfrac{1}{3}$
2. Using conditional independence (standardized)

   $$\text{Prob}(A)=\frac{4}{16}+\frac{1}{16}+\frac{5}{16}=\frac{5}{8}\Rightarrow a=\frac{3}{5};\ \ \text{Prob}(A\,|\,B)=\frac{\frac{1}{4}+\frac{1}{16}}{\frac{1}{4}+\frac{1}{16}+\frac{1}{8}+\frac{1}{4}}=\frac{5}{11}\Rightarrow b=\frac{6}{5}$$

   $$\text{Prob}(A\,|\,C)=\frac{\frac{5}{16}+\frac{1}{16}}{\frac{5}{16}+\frac{1}{16}+\frac{1}{8}}=\frac{3}{4}\Rightarrow c=\frac{1}{3}$$

   Applying Eq. (4) with τ=1: $x=\dfrac{2}{3}\Rightarrow \text{Prob}(A\,|\,B,C)=\dfrac{3}{5}$
3. Using Monte Carlo simulation on the probability perturbation algorithm (with τ=1 in Eq. (4), $\text{Prob}(A\,|\,B,C)=0.35$
4. Using Monte Carlo simulation on the “gradual deformation of sequential simulation” (Hu et al., 2001) $\text{Prob}(A\,|\,B,C)=0.27$

It is clear from comparing [1.] and [2.] that the conditional independence hypothesis is not valid for this case.

While the PPM relies on the same assumption of conditional independence the result is much closer to the true posterior probability. The reason for the latter observation can be explained by means of Eq (5). In this equation, the pre-posterior Prob($I(\mathbf{u}_j)$=1|C) is a function of the data C through the parameter r_C, *and*, a function of an initial realization

$\{i(\mathbf{u}_1),i(\mathbf{u}_2)=1,i(\mathbf{u}_3)\}$. This initial realization depends itself on the pre-posterior Prob($I(\mathbf{u}_j)=1|\ B_j$) with B_j depending on the random path taken. Hence, Eq. (5) forces an explicit dependency between the Prob($I(\mathbf{u}_j)=1|B_j$) and Prob($I(\mathbf{u}_j)=1|$C) *prior* to combining both into Prob($I(\mathbf{u}_j)=1|\ B_j$,C) using a conditional independence hypothesis (Eq. (4) with $\tau=1$). At least from this simple example, one can conclude that the *sequential* decomposition of the posterior into pre-posteriors has robustified the estimate of the true posterior under the conditional independence hypothesis.

The same conclusion can be reached for the gradual deformation of sequential simulation. In gradual deformation of sequential simulation one perturbs gradually the random numbers used to draw from the various conditional distributions in Eq.(2), not the conditional distributions themselves as in the probability perturbation method. It appears that the gradual deformation of sequential simulation has an implicit model of dependency between the B and C data different from the probability perturbation, and more importantly different from the actual dependence.

The differences between the various methods are considerable. One can therefore conclude that future research should focus on understanding better the basic model assumptions, such as conditional independence, rather than focussing on developing precise samplers of models that are based on poorly understood assumptions. Such assumptions will have a first order effect on the ultimate space of uncertainty created.

References

Besag, J, and Green, P.J.,. Spatial statistics and Bayesian Computation, *Journal of the Royal Statistical Society*, B, v. 55, 1993, p. 3-23.

Caers, J., History matching under a training image-based geological model constraint. *SPE Journal*, SPE # 74716, 2003, p. 218-226

Hoffman, B.T. and Caers, J., Geostatistical history matching using the regional probability perturbation method. *In* SPE Annual Conference and Technical Exhibition, Denver, Oct. 5-8. SPE # 84409, 2003, 16pp. Society of Petroleum Engineers.

Hoffman, B.T. and Caers, J., History matching with the regional probability perturbation method – applications to a North Sea reservoir *In*: Proceedings to the ECMOR IX, Cannes, Aug 29 - Sept 2, 2004.

Hu, L.Y, Blanc, G. and Noetinger, B., Gradual deformation and iterative calibration of sequential stochastic simulations**.** *Mathematical Geology*., v 33, 2001, p. 475-490.

Journel, A.G., Geostatistics: roadblocks and challenges. In Soares, A. ed., Geostatistics-Troia, v1: Kluwer Academic, Dordrecht, 1993, p. 213-224.

Journel, A.G., Combining knowledge from diverse data sources: an alternative to traditional data independence hypothesis. *Mathematical. Geology*., v. 34, 2002, p. 573-596.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. and Teller, E., Equation of state calculation by fast computing machines. *J. Chem. Phys*., v. 21, 1953, p. 1087-1092.

Moosegard, K. and Tarantola, A., Monte Carlo sampling of solutions to inverse problems. *Journal of Geophyical. Research, B*, v. 100, 1995, p. 12431-12447.

Neal, R.M., Probabilistic inference using Markov chain Monte Carlo methods. Technical report CRG-TR-93-1, 1993 Department of Computer Science, University of Toronto.

Omre, H. and Tjelmeland, H. Petroleum Geostatistics. *In* Proceeding of the Fifth International Geostatistics Congress, ed. E.Y Baafi and N.A. Schofield, Wollongong Australia, v.1, 1997, p. 41-52. Kluwer Academic Publishers.

Strebelle, S., Conditional simulation of complex geological structures using multiple-point geostatistics. *Mathematical Geol*ogy, v. 34, 2002, p. 1-22.

SPECTRAL COMPONENT GEOLOGIC MODELING: A NEW TECHNOLOGY FOR INTEGRATING SEISMIC INFORMATION AT THE CORRECT SCALE

TINGTING YAO, CRAIG CALVERT, GLEN BISHOP,
TOM JONES, YUAN MA, LINCOLN FOREMAN
ExxonMobil Upstream Research Company,
Houston, TX 77252, U.S.A.

Abstract. Spectral component geologic modelling (SCGM) is a new technology developed to properly account for both the scale and accuracy of any and all interpretations derived from seismic data in building geologic models. Seismic data can be integrated as spectral components which are volume- or map-based property interpretations representing a specific and measurable scale. The SCGM method starts with combining different spectral components together to build what is referred to as a "tentative geologic model", accounting for different scales and measurement accuracy of information in each component. The tentative geologic model will then be further constrained to honor the spatial continuity by substituting the amplitude spectrum of current tentative model with the desired amplitude spectrum from the target variogram model through spectral simulation. After that, the model will then be post processed to first honor the target histogram and then well data. In addition to honoring one single global variogram model, as do traditional geostatistical algorithms, SCGM has the capability to model local variations or trends in the continuity range and dominant azimuth direction of spatial continuity, by modifying the amplitude spectrum using spectral simulation.

1 Introduction

Geologic modeling has been widely used in reservoir management to characterize the rock-property heterogeneity that control pore-fluid storage and flow in a reservoir. For many reservoirs, particularly those in discovery through early production stages, well data may be sparse, and the well data alone are often insufficient to adequately constrain the assignment of reservoir properties between the wells in the geologic model. For such reservoirs, 3D seismic data have been increasingly used as an aid to assign these properties in the geologic model.

However, the utilization of seismic data for modeling reservoir properties faces some severe problems, possibly the most important being that of the difference in scale and accuracy between the seismic and the well data. The traditionally used geostatistical modeling methods integrate seismic data through kriging with local varying mean (Goovaerts, 1997), block cokriging (Behrens, Macleod, and Tran, 1996; Yao, 2000), or simulated annealing (Deutsch, Srinivasan, and Mo, 1996). These methods either

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 23-33.

wrongly treat the seismic data at the same scale as the geologic model, or make some strong assumption about the relationship between the coarse-scale seismic data and fine-scale well data (linear average as in block co-kriging). As a result, the geologic model may not fully exploit information contained in the seismic data and may not honor the input information.

The primary incentive for developing the SCGM method was to properly account for both the limitation of scale and accuracy of any and all interpretations derived from seismic data. Secondary incentives were to obtain a method that provides advanced capabilities for controlling rock-property continuity in the geologic model, and that can build or truly update a geologic model with new information quickly, based on spectral simulation (Calvert et. al., 2000, 2001, 2002).

2 Review of spectral simulation

Spectral simulation is gaining wider application in building geologic models due to the advantage of better honoring the spatial continuity of petrophysical properties, such as reservoir property and shale volume. The spatial continuity structure is characterized by a covariance/variogram model in the space domain and is represented by a density spectrum in the frequency domain. Distinct from sequential simulation methods, spectral simulation is a global method in the sense that a global density spectrum is calculated once from variogram model and the inverse Fourier transform is performed on the Fourier coefficient only once to generate a realization.

A spectral-simulation method, called Fourier Integral Method (FIM), has been proposed to generate geologic-model realizations that honor the spatial structure of a random field $z(u)$ in one-, two-, or three dimensions (Borgman, Taheri, and Hagan, 1984; Gutjahr, Kallay, and Wilson, 1987; Mckay, 1988; Pardo-Iguzquiza and Chica-Olmo, 1993). This method is performed in the frequency domain, as opposed to the usual sequential-Gaussian-simulation method performed in the space domain.

The spatial structure of a random field $z(u)$ is characterized by the covariance $C_z(h)$ or variogram $\gamma_z(h)$ in the space domain. In 1D, the covariance of $z(u)$ is defined as the convolution product (Bracewell, 1986):

$$C_z(h) = \int_{-\infty}^{+\infty} z(u) \cdot z(u+h)du = z * \overset{\vee}{z}\text{, where } \overset{\vee}{z}(u) = z(-u) \qquad (1)$$

The Fourier transform (FT) of the covariance into the density spectrum of $z(u)$ in the frequency domain exchanges convolution and multiplicative products:

$$s(\varpi) = FT(C_z) = FT(z) \cdot FT(\overset{\vee}{z}) = Z(\varpi) \cdot Z^*(\varpi) = | Z(\varpi) |^2 \qquad (2)$$

where $Z(\varpi) = FT(z) = \int z(u)e^{-i\varpi u}du$, and $Z^*(\varpi)$ is the complex conjugate. The term $s(\varpi)$ is referred to as the density spectrum, $|Z(\varpi)|$ as the amplitude spectrum, $Z(\varpi) = |Z(\varpi)|e^{-i\phi(\varpi)}$ as the Fourier coefficient, and $\phi(\varpi)$ as the phase. The spectral-simulation method is based on the correspondence between the space property, $z(u)$, and the frequency counterparts, $s(\varpi)$ and $\varphi(\varpi)$, as illustrated in Figure 1. The implementation details can be referred to Yao (1998, 2002)

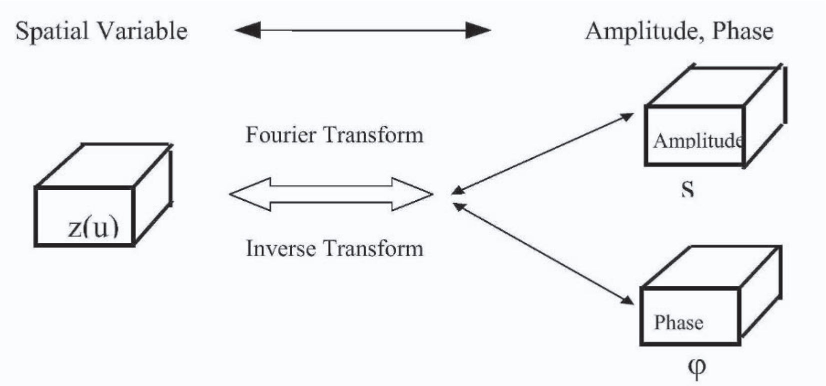


Figure 1. The correspondence between the space domain variable, z, and the frequency counterparts, $s(\varpi)$ and $\varphi(\varpi)$.

There are several advantages of spectral simulation over traditional geostatistical simulation. The spectral-simulation method is fast, particularly when based on the Fast Fourier Transform (Kar, 1994; Lam, 1995; Bruguera, 1996; Mckay, 1998). It is a global method in the sense that all of the amplitude-spectrum values over the whole field are used simultaneously to generate the simulated property. Therefore, the amplitude spectrum, or variogram model in the space domain, can be honored globally over the whole field instead of only within search neighborhoods as with the traditional sequential-Gaussian simulation method. Actually, the variogram model is honored over half field size, see Yao, 2002. A related advantage is that the separation of amplitude (spatial continuity) and phase (spatial location) allows updating of models if new information about either spatial continuity or location is obtained, or allows the conditioning of models to local information; see Calvert et al (2001, 2002). In addition, the advantage of spectral simulation in separating amplitude and phase information allows the spatial continuity to be modified to account for this traditionally unaccountable local information (Calvert et al., 2001, 2002).

3 Scale and accuracy of interpreted rock-property information: spectral component

All data that we use for geologic interpretation are limited in the scale of rock-property information that they contain, although we do not always appreciate the fact. For example, seismic data cannot directly be used to predict high-frequency variability in rock properties because the seismic data contain no information at high frequencies. If we attempt to use seismic data to estimate rock properties at scales that are outside of the data frequency band, interpretation errors can result. A spectral component is a volume- or map-based property interpretation representing only a specific and measurable scale of the property to be modeled, such as porosity. It could be derived from any data source or even from analogue information, representing all or a portion of the reservoir volume being modeled. The following data represent some of the spectral components (Figure 2):

- 3D volume from seismic amplitude calibration, which contains information only within the seismic frequency band, typically about 15-75 Hz.
- 2D map from seismic facies or geologic interpretation, such as average porosity map. This provides no information about the vertical variability in porosity values, hence contains no information at any frequency above zero Hz in vertical direction.
- 1D trend from well data, such as compaction trend of porosity observed, i.e., pososity generally decreasing with burial depth according to a fairly predictable function. This contains only low frequency information, e.g., 0-5 Hz, because slow vertical changes represent low-frequency vertical variability.

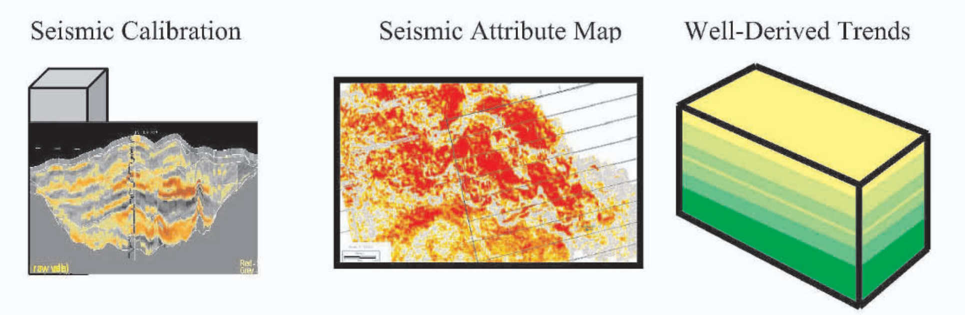


Figure 2. Examples of spectral components at different scales.

These spectral components are generated from different sources - some may be generated from data interpretation, whereas others may be generated from a concept or an analogue. Different frequency components might have different accuracy - those directly measured from well logs will be more accurate than others from qualitative interpretation. In addition, a spectral component often contains information that spans a bandwidth of frequencies. The measurement/interpretation accuracy could also change with different frequency component. The new SCGM method will account for the uncertainty or accuracy about each spectral component.

4 Overview of SCGM method

SCGM method involves constructing a geologic model by first mathematically combining different spectral components together. The combined volume is referred to as the "tentative geologic model", which might not represent all desired reservoir characteristics and needs to be further constrained to honor the target statistics such as variogram, histogram and well data. The constrained model can be further post processed to represent local variations in continuity trend and continuity azimuth, based on spectral simulation. The generalized SCGM process is given in Figure 3.

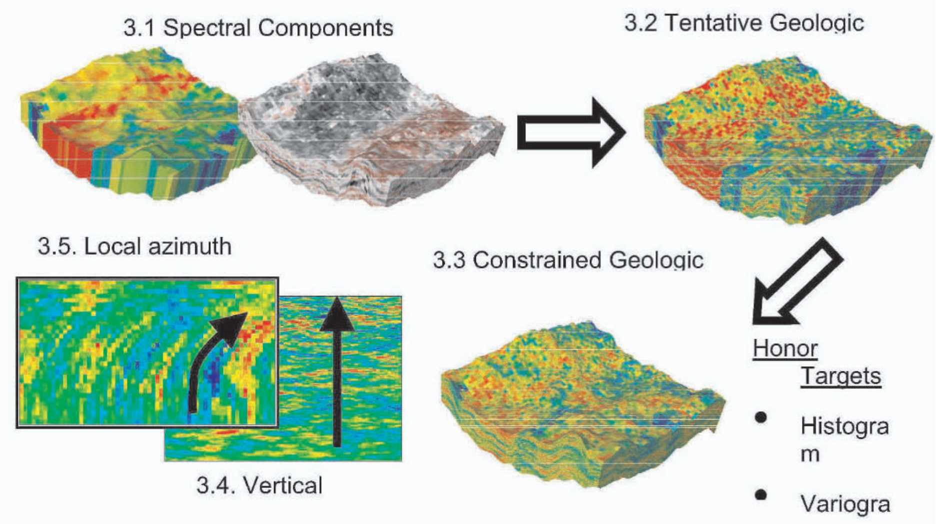


Figure 3. General process of SCGM method.

4.1 BUILDING THE TENTATIVE GEOLOGIC MODEL

The SCGM method starts with combining different spectral components together to build what is referred to as a "tentative geologic model", accounting for different scales of information in each component and the different measurement accuracy. The tentative geologic model represents the integration of all relevant data types to produce an *a priori* geologic model. The least-complex tentative geologic model that can be built is one in which all spectral components represent distinct or complementary frequency bands. Such a tentative geologic model can be constructed by simply summing the independent components. However, in reality, there will always be missing scales of information. For any tentative geologic model that is built, if a frequency-band of information is missing (e.g., high-frequency information), then this missing band of information must be simulated within SCGM and added to the tentative geologic model. It is also possible that information may be missing over a specific region within the model area; in this case the data are simulated and added to the model, but only in that specific region.

The process of building the tentative geologic model gets somewhat more complicated when the individual spectral components overlap in their frequency content. The spectral components may completely overlap in frequency content or they may partially overlap. To properly integrate spectral components that have overlapping spectra, the measurement or interpretation accuracy of each overlapping component must be known. The accuracy can be quantified by a value between 0.0 and 1.0. Those components having higher accuracy values will have relatively greater influence on the resulting tentative geologic model, but only over those frequencies that overlap. Measurement or interpretation accuracy can also vary spatially. For example, the spectral component in one location within the modeling area may be more accurate than in another due to the interpreter's diligence. In this case, the tentative geologic model is constructed through weighted averaging of the spectral components *by location* (i.e., different average weight at different locations). Figure 3.2 shows a simple tentative model generated by simply adding up different frequency components in Figure 3.1.

4.2 CONSTRAINING THE TENTATIVE GEOLOGIC MODELS

The tentative geologic model will not have all of the desired properties of geologic model, e.g., it likely will not honor the well data or the target variogram and histogram. The tentative geologic model is further constrained to honor these targets. The constraining process is sequential, in that the model is modified first to honor the variogram, second to honor the histogram, and finally to honor the well data.

- *Honoring the desired spatial continuity.* Spectral simulation is a perfect application to update the tentative geologic model in an attempt to honor the spatial continuity represented by the target variogram model. From the tentative geologic model, we calculate its amplitude and phase spectrum. We only keep part of the amplitude spectrum which we believe are reliable and substitute the other part with the target one (representing the target variogram model), and keep the phase spectrum to generate new Fourier coefficients. The inverse Fourier transform provides a model that honors the spatial continuity represented by the target amplitude spectrum, as well as the spatial distribution of high and low values observed in the tentative geologic model, see Figure 4.

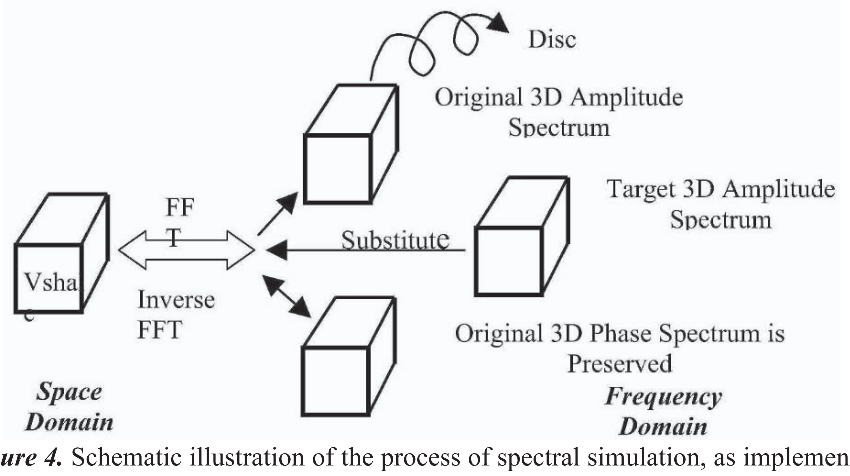

Disc
Original 3D Amplitude Spectrum
Target 3D Amplitude Spectrum
FF T
Substitute
Vsha e
Inverse FFT
Original 3D Phase Spectrum is Preserved
Space Domain
Frequency Domain

Figure 4. Schematic illustration of the process of spectral simulation, as implemented in SCGM.

- *Honoring the desired histogram***.** Quartile transform is used to force the distribution of geologic model matches the target histogram. Each rock-property value in each cumulative distribution function (CDF) corresponds to a probability quintile. The p quintile of the cumulative distribution function (CDF) is transformed to the same p quintile of the CDF for the target histogram so that the CDF of the tentative geologic model matches the target CDF. The quintile transform does not change the relative rank of the data, hence also referred to as "rank-preserved transform".

- *Honoring the well data.* Following the steps of honoring the variogram model and the histogram, the rock-property values at the well locations are reset to match the actual well-data values. Resetting these values results in changes in the properties assigned to these cells. These changes [(actual well value) - (current model value)] are propagated to all tentative geologic model cells in the neighborhood of each well; the magnitudes of the changes are weighted as a function of inverse distance from the well.

Figure 3.3 shows a constraint model which honors the target variogram, histogram, and well data. Note that the sequential implementation of first honoring variogram, then honoring histogram, finally honoring well data might distort the target parameters honored first such as varigoram and histogram by later honoring other target parameters such as well data. Therefore, the ideal implementation would be iteratively repeat the sequential honoring process to ensure all the target parameters are honored in the same degree. However, many tests show that so long as the target parameters are consistent with each other, the first iteration does 90% of the job of constraining the model to the target. For speed purpose, we used only one iteration, but strongly suggest checking the model to make sure all the targets are met without much distortion.

The SCGM process described above *does* properly account for the scale of the input data, both in terms of the spectral component information (as represented in the phase spectrum) and the target variogram model (as represented in the target amplitude spectrum). As a result, SCGM can honor both the compositional information contained in the tentative geologic model and the target variogram model, without compromising either. Given the same input of variogram, histogram and well data, SCGM is proved to honor the input information better than the traditional geostatistical methods such as kriging with locally varying mean or collocated cokriging (Calvert et al., 2002).

In addition to honoring one single global variogram model, as do traditional geostatistical algorithms, SCGM has the capability to model local variations or trends in the range and dominant azimuth direction of spatial continuity, using spectral simulation.

4.3 CONDITIONING TO THE LOCAL SPATIAL CONTINUITY TREND

In a geologic model, the three-dimensional spatial continuity of a rock property is commonly controlled with geostatistical algorithms and a variogram that quantifies the spatial variability of the rock property as a function of both separation distance and direction. Geostatistical algorithms used in constructing geologic models assume stationarity in the geologic characteristics of the modeled region, i.e., they assume that a modeled rock property can be represented by a single set of statistical measures, which are often referred to as "global" measures. For example, a single, global variogram model would be used to represent the spatial continuity of the rock property over the entire modeled region.

However, we know that the geologic characteristics of the subsurface are non-stationary. For example, the global spatial continuity of bed thickness in a reservoir can be characterized by a global spherical variogram model with a range of 10 feet. However, we can also observe a trend of thicker beds at the bottom and thinner beds on the top: the thicker beds at the bottom might have a range of 20 feet and the thinner beds on the top might have a range of 5 feet. To account for the local continuity trend beyond the global variogram model, we use the local (variogram models with longest and shortest ranges) and global variogram models to calculate the spectral amplitude ratios of the local spectral amplitudes vs. the global one at different frequency bins. We interpolate the local amplitude ratios in between according to the trend. This will provide one local amplitude ratio for each frequency bin at each cell. Then, we decompose the model that honors the global spectrum into different frequency components (represent each frequency bin), and multiply each component by the corresponding local amplitude ratio. The summed result of all the multiplied spectral components displays the local continuity trend. The implementation details can be referred to Yao (2003).

Figure 3.4 gives examples of applying this method to impose a trend in vertical continuity on a geologic model of shale volume. This example applied only one continuity trend on the geologic model. If additional trends in spatial continuity are desired, then treat the geologic model created before as a new starting geologic model and apply the local trend using different global and local amplitude spectra that represent the new trends. This will generate a geologic model that honors multi-dimensional trends in spatial continuity.

Other algorithms available to account for non-stationary spatial continuity usually separate the whole modeling area into different sub-areas and use a different variogram model for each sub-area. Such a method addresses the non-stationarity of the large area, but at the cost of artifact boundaries between sub-areas. Using spectral simulation and manipulating the amplitude spectrum allows us to address explicitly the gradually changing continuity trend.

4.4 CONDITIONING TO THE LOCAL SPATIAL CONTINUITY AZIMUTHS

Rock property continuity within a reservoir often shows anisotropy, i.e., continuity is greater in one direction than in another. In addition, the local direction of greatest continuity might change from one location to another within the reservoir. Consider sediments deposited in a river channel. Paleo-hydrodynamics often control the distribution of the lithological and petrophysical properties within the channel. We know that the continuity of these properties is anisotropic, typically greatest along channel and less continuous across channel. We also know that sinuosity may cause the channel to locally vary in direction; therefore, the rock-property continuity will also locally vary in direction, as with a meandering pattern.

Methods used to introduce variable directions in rock-property continuity into the geologic model are generally based on methods published by Xu (1996). Continuity direction is varied according to an input grid of azimuths, which represent local variations in continuity direction. Often, this grid is based on interpreted seismic facies, i.e., the shape of the interpreted facies (e.g., sinuosity of a channel) is represented in the azimuth data. To conditioning to the locally varying continuity azimuth, we first identify the strings of connected nodes from the azimuth grids (Jones, et al., 2001). Then, we simulate each string to have maximum continuity along that string, using 1D spectral simulation. Finally, we put the simulated values back to the original nodes along the string. Therefore, the continuity along a path that bends according to the azimuth data as desired is reproduced (Craig, et al., 2002). The traditional pixel-based geostatistical methods require that this path be represented locally by a straight line. If those segments are small (i.e., the range of continuity is short), the curved line can be approximated well with straight-line segments. However, if the segments are long (i.e., the range of continuity is long), the curved line can not be approximated with straight-line segments. This limitation practically manifests itself as a trade-off between honoring the azimuth data and the target variogram range. The new SCGM method can simulate continuous rock properties along a bent path, therefore, it should produce better results in situations when long-range continuity is to be represented along a curved geologic feature, see Figure 3.5.

4.5 UPDATING AN EXISTING GEOLOGIC MODEL

The process of updating of any existing geologic model with any new information is very straightforward. For example, a new or alternative spectral component (e.g., from seismic-volume interpretation) became available after the original model was built. To incorporate this new information, we could update the original model by the following process:

- From the existing model, filter out and discard the information that is of the same scale (frequency band) as the new spectral component,
- Combine this filtered model with the new spectral component to create a tentative geologic model
- Constrain the new tentative geologic model to satisfy other targets.

A new variogram or histogram target model can also be incorporated efficiently to update the existing model.

5 Conclusions

SCGM is a new technology for integrating all the relevant data at their correct scale. It starts with combining different spectral components together to build what is referred to as a "tentative geologic model", accounting for different scales of information in each component and the different measurement accuracy. Then, the tentative geologic model is constrained to honor all the desired properties of geologic model such as honoring the spatial continuity, histogram and well data. Spectral simulation is applied to honor the spatial continuity globally as well as to gain speed advantage. In addition to honoring one single global variogram model, as do traditional geostatistical algorithms, SCGM has the capability to model local variations or trends in the range and dominant azimuth direction of spatial continuity, by modifying the amplitude spectrum using spectral simulation.

References

Behrens, R.A., Macleod, M.K., and Tran, T.T., 1996. Incorporating seismic attribute maps in 3-D reservoir models: SPE 36499, 31-36.

Borgman, L., Taheri, M., and Hagan, R., 1984. Three-dimentional frequency-domain simulations of geological variables, in Verly, G., Journel, A.G., and Marechal, A., eds., Geostatistics for natural resources characterization: NATO ASI Series, Reidel Publ., Dordrecht, The Netherlands, p. 517-541.

Bracewell, R., 1986. The Fourier transform and its application: McGraw Hill, Inc., Singapore, 474 p.

Bruguera, J., 1996. Implementation of the FFT butterfly with redundant arithmetic: IEEE Transaction on Circuits and Systems, Part II: Analog and Digital Signal Processing, v. 43, no. 10, p. 717-723.

Calvert, C.S., Bishop, G.W., Ma, Y.Z., Yao, T., Foreman, J.L., Sullivan, K.B., Dawson, D.C., and Jones, T.A., 2000. Method for constructing 3D geologic models by combining multiple frequency pass band, U.S. Patent Application Pending.

Calvert, C.S., Yao, T., Bishop, G.W., and Ma, Y.Z., 2001. Method for locally controlling spatial continuity in geologic models, U.S. Patent Application Pending.

Calvert, C.S., Jones, T.A., Bishop, G.W., Yao, T., Foreman, J.L., and Ma, Y.Z., 2002. Method for conditioning a random field to have directionally varying anisotropic continuity, U.S. Patent Application Pending.

Calvert, C.S., Bishop, G.W., Yao, T., Ma, Y., Forema, J.L., 2002. Spectral Component Geologic Modeling: an overview, ExxonMobil internal report.

Deutsch, C.V., Srinivasan, S., and Mo, Y., 1996. Geostatistical reservoir modeling accounting for precision and scale of seismic data: SPE 36497, 9-15.

Goovaert, P., 1997. Geostatistics for natural resources evaluation, Oxford Univ. Press.

Gutjahr, A., Kallay, P., and Wilson, J., 1987. Stochastic models for two-phase flow: A spectral-perturbation approach: Eos, Transaction, American Geophysical Union, v. 68, no. 44, p. 1266-1267.

Jones, T.A., Foreman, J.L., Yao, T., 2001. Method to honor channel directionality when building 3-D petrophysical models, U.S. Patent Application Pending.

Kar, D., 1994. On the prime factor decomposition algorithm for the discrete sine transform: IEEE Transactions on Signal Processing, v. 42, p. 3258-3260.

Lam, K.-M., 1995. Computing the inverse DFT with the in-place, in-order prime factor FFT algorithm: IEEE Transactions on Signal Processing, v. 43, p. 2193-2194.

Mckay, D., 1988. A fast Fourier transform method for generation of random fields: Unpubl. Master's thesis, New Mexico Institute of Mining and Technology, 92 p.

Pardo-Iguzquiza, E., and Chica-Olmo, M., 1993. The Fourier integral method: An efficient spectral method for simulation of random fields: Math. Geology, v. 25, no. 4, p. 177-217.

Xu, W., 1996. Conditional curvilinear stochastic simulation using pixel-based algorithms: Math. Geology, v.28, no. 7., p. 937-949.

Yao, T., 1998. Conditional spectral simulation with phase identification: Math. Geology, v. 30, no. 3, p. 285-308.

Yao, T., 1998. SPECSIM: A fortran-77 program for conditional spectral simulation in 3D: Computer & Geosciences, v. 24, no. 10, p. 911-921.

Yao, T., 1998. Automatic modeling of (cross) covariance tables using fast Fourier transform: Math. Geology, v. 30, no. 6., p. 589-615.

Yao, T., and Journel, A.G., 2000. Integrating seismic attribute maps and well logs for porosity modeling in a west Texas carbonate reservoir: addressing the scale and precision problem: Journal of Petroleum Science and Engineering, v. 28, p. 65-79.

Yao, T., 2002. Reproduction of mean, variance, and variogram model in spectral simulation: Math. Geology, accepted.

Yao, T., Calvert, C., Jones, T., Bishop, G., Ma, Y., Foreman, L., 2003. Spectral simulation and its advanced capability of conditioning to local continuity trends in geologic modeling: Journal of Petroleum Science and Engineering, submitted.

Autobiography:

Tingting Yao, BS from University of Petroleum in 1993, majoring in petroleum geology; PhD in Geostatistics from Stanford University in 1998. Her PhD research focuses on automatic covariance modeling to characterize the spatial continuity of earth science phenomena and conditional spectral simulation through phase identification to honor the spatially sampled conditioning data. She Joined Mobil Technology Company in July 1998 and transferred to ExxonMobil Upstream Research Company in February 2000. She has been working on improved methods for more accurate and efficient geologic modeling and reservoir characterization through integration of various data type.

This page intentionally blank

JOINT SIMULATIONS, OPTIMAL DRILLHOLE SPACING AND THE ROLE OF THE STOCKPILE

A. BOUCHER, R. DIMITRAKOPOULOS and J.A. VARGAS-GUZMAN
WH Bryan Mining Geology Research Centre
The University of Queensland, Brisbane Qld 4072, Australia

Abstract. Infill and grade control drilling are a major cost in any mining operation. Reduction of drilling density can considerably enhance the profitability of an operation provided the cost from block misclassification is less than the savings in drilling. This paper presents a general simulation based approach to assess the performance of potential drilling schemes from the available deposit information. The approach integrates joint simulation of correlated variables with the computationally efficient minimum/maximum autocorrelation factors, multi-elements ore classification, and mine planning considerations. The latter employs key indicators such as profit per tonne mined and profit per tonne milled, as well as the potential use of a stockpile and its discounting. A case study at the Murrin Murrin nickel-cobalt deposit, Western Australia, is used to elucidate the proposed approach and to show the critical effect of planning decisions on drilling.

1 Introduction

Infill drilling is a critical information collection process in mining operations leading to substantial investment that can be in the order of millions of dollars. As a result, the ability to assess the performance of potential drilling schemes, prior to drilling is important. A reduction in drilling density could enhance the profitability of an operation, if misclassification cost does not exceed the saving in drilling. At the same time, additional information becomes counterproductive after the point of diminishing returns, i.e. the cost of additional information exceeds its benefit. Past work in geostatistically assessing additional drilling was based on estimation variances which largely reflect the geometry of drilling configurations (Goovaerts, 1997) without any consideration of the local grade variability and uncertainty (Ravenscroft, 1992), economic cost/benefit analysis, or a link to mine planning decisions.

A stochastic simulation framework can be used to realistically address the assessment of infill drilling patterns (e.g., Dimitrakopoulos, 2003). In the general case of multi-element deposits, joint simulation of pertinent correlated variables is used to produce realisations of an exhaustively known deposit. Such a realisation is treated as an "actual" deposit and is subsequently virtually drilled. This new drilling information can then be used to re-simulate the deposit leading to comparisons of block classifications

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 35-44.

and other indicators to the actual exhaustively known deposit. This way different drilling schemes can be compared to make informed selection of a drilling strategy.

Computationally efficient joint simulation methods are essential for generating realistic representation of "actual" deposit. Conventional co-simulation methods (e.g., Verly, 1993; Chiles and Delfiner, 2000) become inefficient when more than two variables are considered. Collocated cosimulation with a Markov-type coregionalisation (Almeida and Journel 1993) assumes a very specific coregionalisation and does not extend well beyond two attributes. Therefore, conditional simulation with the so-called minimum/maximum autocorrelation factors or MAF (Switzer and Green 1984; Desbarats and Dimitrakopoulos 2000) is advocated herein. MAF transforms attributes of interest to uncorrelated factors that are independently simulated by any simulation method and then reconstructed to realisations of the original variables reproducing their cross and auto-correlation.

In addition to the orebody geology, mine planning aspects, such as stockpiling, also affect the performance of the infill drilling patterns,. When low-grade ore blocks are stockpiled, the performance of an infill drilling scheme is a function of how, when, and if the stockpile would be processed in the future. The uncertainty linked to the stockpiling strategy can be factored into the selection of a drilling scheme by depreciating the stockpile value with a discount rate. If an ore block sent to the stockpile is considered lost, the stockpile can be regarded as waste. Alternatively, if the stockpile will be processed in coming years, a misclassified ore block in the stockpile makes no difference and no penalty is necessary. Discount rates enable the comparison of the different schemes by linking the two extreme scenarios.

In the next sections, the drilling optimisation method is outlined and is followed by a brief discussion of simulation with MAF and the definition of economic indicators for comparing drilling efficiency. Finally, the intricacies of the method are detailed in a case study at the Murrin Murrin nickel-cobalt deposit, Western Australia.

2 A method for infill drilling assessment and optimisation

The following method, also schematized in Figure 1, is suggested to assess and compare the performance of drilling patterns for a multi-element deposit:

Step 1: From the exploration drilling data available within the pit, jointly simulate a representation of the deposit using min/max autocorrelation factors for the attributes under study. This first realization is called the "actual" deposit.

Step 2: Sample the above actual deposit with the different infill drilling schemes of interest.

Step 3: For each drilling scheme, jointly simulate with MAF the elements of interest conditional to the data from the "actual" deposit in Step 2 above, to obtain several joint realisations. Re-block the realizations for the attributes simulated to produce models of mining blocks to be assessed.

Step 4: Do grade control and classify the blocks (e.g. from their average grades) for each sampling scheme (e.g., milling ore, stockpile and waste). Compare the classification to the "actual" classification using, as economic indicators, the profit per tonne mined and profit per tonne milled; calculated as discussed in a subsequent paragraph.

Step 5: Graph and assess the results with respect to the point of diminishing returns. Repeat to assess sensitivity of the results.

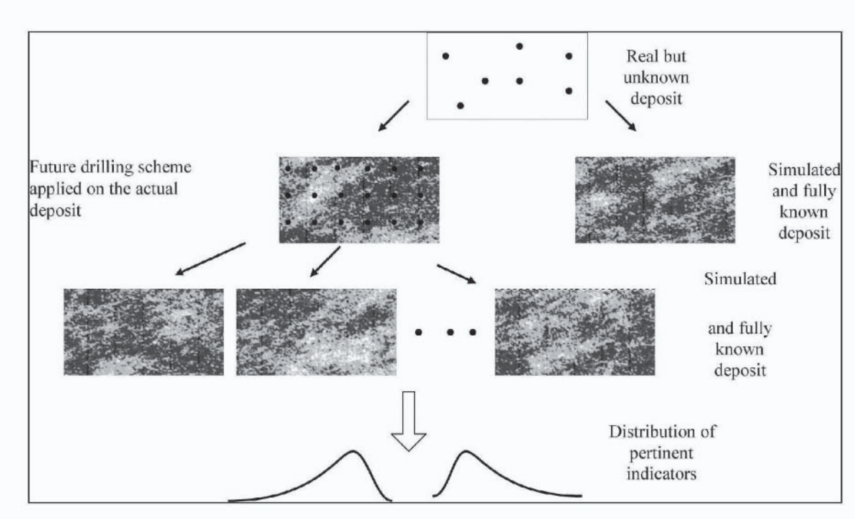

Real but unknown deposit
Future drilling scheme applied on the actual deposit
Simulated and fully known deposit
Simulated
and fully known deposit
Distribution of pertinent indicators

Figure 1 Schematic workflow for the proposed methodology

3 Multivariate simulation with minimum/maximum autocorrelation factors

The multivariate deposit is cosimulated by orthonalising the grade attributes into three factors deemed uncorrelated from each other. Each of these factor is then simulated independently, and the simulated grade is obtained by back-rotating the factors into the attribute space.

The minimum/maximum autocorrelation factors (MAF) is an orthogonalisation similar to the well-known principal components analysis (PCA). The advantage of the MAFs over the PCs is the extension of orthogonalisation to the non-zero lags. Principal components (David, 1988) are uncorrelated at all lags only if they are derived from a random field with an intrinsic coregionalisation. The MAFs uncorrelate a RF for all lags with a linear model of coregionalisation containing at most two structures (Switzer and Green 1984; Desbarats and Dimitrakopoulos 2000). Boucher (2003) and Dimitrakopoulos and Fonseca (2003) have successfully used the MAF method to jointly simulate grades in a mining environment.

Switzer and Green (1984), later reviewed in Desbarats and Dimitrakopoulos (2000), show how to obtain the factors through the eigenvectors of the matrix $2\mathbf{\Gamma}_Z(h)\mathbf{B}^{-1}$, where

$$\mathbf{B} = \text{cov}[\ \mathbf{Z}(u)\ ,\ \mathbf{Z}(u)\ ]$$
$$2\mathbf{\Gamma}_Z(h) = \text{cov}[\ \mathbf{Z}(u) - \mathbf{Z}(u+h),\ \mathbf{Z}(u) - \mathbf{Z}(u+h)\ ]$$

where $\mathbf{B}$ is the variance/covariance matrix of $\mathbf{Z}(u)$, a multiGaussian RF, and $\mathbf{\Gamma}_Z(h)$ is the variogram matrix at lag *h*.

The matrix $\mathbf{A}$ of orthogonalisation coefficients is such that

$$2\mathbf{\Gamma}_Z(h)\mathbf{B}^{-1} = \mathbf{A}^T \Lambda \mathbf{A} \tag{1}$$

Refer to Desbarats and Dimitrakopoulos (2000) for an equivalent but computationally more efficient method to derive the coefficients $\mathbf{A}$ by performing two successive principal component decompositions.

Each orthogonal factor $\mathbf{Y}_i(u), i=1,...,p$ is obtained with the coefficients $\mathbf{a}_i$ constituting the i*th* row of $\mathbf{A}$.

$$\mathbf{Y}_i(u) = \mathbf{a}_i\,\mathbf{Z}(u),\ \ i=1,...,p \tag{2}$$

The new vector RF $\mathbf{Y}(u)$, is then, by construction orthogonalised at lag 0 and at lag h. $\mathbf{Z}(u)$ is simulated by independently simulating each factor $y_i^*(u), i=1,...,p$ and back-rotating them with the coefficient matrix:

$$\mathbf{z}^*(u) = \mathbf{A}^T \cdot \mathbf{y}^*(u) \tag{3}$$

In practice, the attributes are first normal score transformed and then rotated into orthogonal factors. This prior transformation reduces the effect of skewed distributions but potential problems may occur if the rank and the Pearson coefficient of correlation of the original attribute differ too much.

4 Economic indicators

The key indicators suggested here are the profit per tonnes mined and profit per tonnes milled. These are defined here, without loss of generality, using the case of a deposit with two revenue generating elements and a third one that adversely affects production. Consider an example of a Ni laterite deposit producing nickel and cobalt where magnesium is a “penalty” element increasing processing costs.

Taking into account the quantity of the penalty element (M^{Mg}) and a penalty factor (f_k^{Mg}) the cost of classifying a block at location $\mathbf{u}$ in category *k*, $C_k^{\mathrm{Total}}(\mathbf{u})$ is expressed as

$$C_k^{\mathrm{Total}}(\mathbf{u}) = \left(C^{\mathrm{Drilling}} + C_k^{\mathrm{Mining}} + C_k^{\mathrm{Processing}}\right) + M^{\mathrm{Mg}}(\mathbf{u}) \cdot f_k^{\mathrm{Mg}}$$

The revenue from a block is expressed as

$$R_k(\mathbf{u}) = \sum_{i=1}^{n} M_k^i(\mathbf{u}) r_k^i p^i$$

where M^i is the quantity of revenue generated by metal i, (e.g., M^{Ni} is the quantity of Nickel), r_k^i is the recovery of metal *i* when classified in category *k* and p^i is the price for attribute *i*.

The gross profit $F_k(\mathbf{u})$ generated by classifying a block at location $\mathbf{u}$ in group k is

$$F_k(\mathbf{u}) = R_k(\mathbf{u}) - C_k^{\mathrm{Total}}(\mathbf{u}) \tag{4}$$

The final classification of a block at location $\mathbf{u}$ is such that it maximizes the gross profit. A block will be classified in group k' such that

$$k' = \arg\max_j F_j(\mathbf{u})$$

The optimal drilling pattern is the one that would maximize the gross profit, such that the sum of all $F_{k'}(\mathbf{u}_j), j = 1,...,N$ is maximal. Excessive infill drilling would increase the cost and insufficient drilling would decrease the revenue.

Two indicators are used to assess the performance of the sampling scheme. The primary one is the profit per tonne mined, which is the sum of the profit generated by the N blocks inside the domain

$$P_{\mathrm{mined}} = \frac{\sum_{j=1}^{N} F_{k'}(\mathbf{u}_j)}{N} \tag{5}$$

which is an indication of the efficiency of the selection. The second indicator is the profit per tonne milled

$$P_{\mathrm{milled}} = \frac{\sum_{j=1}^{N} F_{k'}(\mathbf{u}_j)}{\sum_{j=1}^{N} I_{\mathrm{milled}}(\mathbf{u}_j)} \tag{6}$$

where $I_{\text{milled}}(\mathbf{u}_j)$ takes the value one if the block located at $\mathbf{u}_j$ is sent to the mill, and takes zero otherwise. The profit per tonne milled indicates the quality of the ore being selected. The difference between the two can be seen with a simple example. Consider a case where some economics material has to be stockpiled to allow only very high grade to the mill. Being profitable, the misclassification of this stockpile material in the mill material will increase the revenue and, the tones mined being constant, the ratio profit per tones mined will also increase. In contrast, the profit per tone milled will decrease as the misclassified material generates less revenue than the high grade material.

The conditional distribution of profit per tonne mined P_{mined} (5) and per tonne milled P_{milled} (6) are computed for each of the N_D sampling scheme Ωj from their respective conditional joint simulations, obtained from Eq. (1) to (3).

$$\Pr\ (P_{\text{mined}} < X \mid \mathbf{z}(u_i),\ \Omega_j,\quad i = 1,\ldots,n)\quad j = 1,\ldots,N_D$$
$$\Pr\ (P_{\text{milled}} < X \mid \mathbf{z}(u_i), \Omega_j,\ i = 1,\ldots,n)\ \ j = 1,\ldots,N_D$$

5 Application at the Murrin Murrin deposit

The Murrin-Murrin nickel-cobalt deposit is located in the Eastern Goldfields Province, Western Australia. It is hosted in weathered peridotites comprising of a ferrugious zone, which is predominantly waste, and two ore bearing horizons, a smectite unit with a transitional boundary to a magnesium enriched saprolite horizon (Jaine, 2003) The Murin Murin operation provides a 4 Mtpa supply to the processing plant that recovers nickel and cobalt. Given the magnesium content of the ore, the response of the mill feed to pressure-acid leaching and the cost of acid consumption is a metallurgical issue.

In the case study that follows, exploration drillholes within the saprolite zone are available from one of the open pits at Murrin Murin. Those holes, approximately gridded on a 50x50 metre spacing, give 263 one metre composites. Grade control infill drilling is typically performed on a 12.5 by 12.5 metre grid, with block size of 15x15 metre. A reduction in drilling would lead to direct saving in pre-mining costs whilst additional information could improve the quality of mill feed, thus reducing contaminant penalties and improving ore selection in addition to improving short term scheduling performance of the mine. The choice of a bench height is also looked at, two and three metre bench thicknesses are considered.

5.1 SIMULATING Ni, Co AND Mg WITH MAF

From the exploration drillholes, 4 actual deposits are simulated on point support: 2 with two metre bench and 2 with three metre bench. After normal score-transform, the nickel, cobalt and magnesium attributes are rotated into MAF space with expression (2). Each of the factors is then simulated independently one from the other. Then rotated back together into the Gaussian space and finally back-transformed into the original space. Figure 2 shows some joint realisations of nickel, cobalt and magnesium of one actual deposit. The cross-variograms, shown in Figure 3, are well-reproduced thus preserving the important spatial relationships between the attributes.

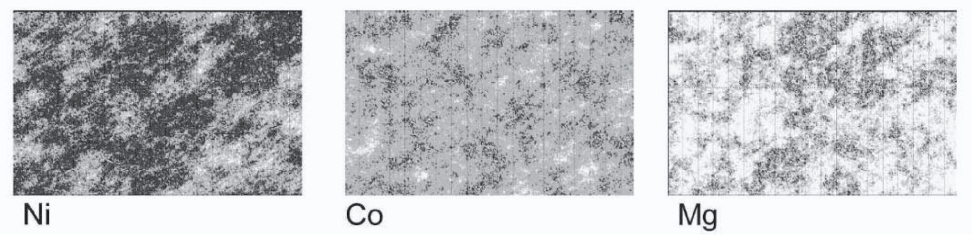


Figure 2 Joint realisations of nickel, cobalt and magnesium. Light is low, dark is high.

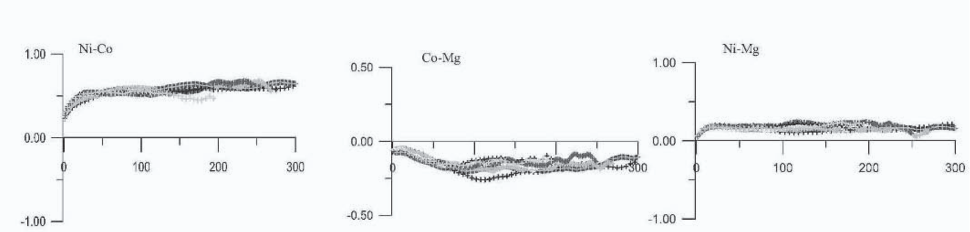


Figure 3 Reproduction of cross-correlation at all lags. The black crosses are the experimental variogram values.

5.2 SIMULATING THE DRILLING AND CLASSIFICATION PROCESS

The infill drilling information $\mathbf{z}(\mathbf{u}_\alpha) = \{z_{Ni}(\mathbf{u}_\alpha), z_{Co}(\mathbf{u}_\alpha), z_{Mg}(\mathbf{u}_\alpha)\}$, where $\mathbf{u}_\alpha$ are the sampling locations, is obtained by virtually drilling the actual deposits with a specific drilling scheme. Four ($N_D = 4$) regular sampling schemes, $\Omega_i, i = 1,...,4$, are considered:

Ω_1 :	12m x 12m	(512 holes)
Ω_2 :	18m x 12m	(320 holes)
Ω_3 :	18m x 18m	(210 holes)
Ω_4 :	25m x 25m	(210 holes)

For each of these sampling, 30 cosimulations $\mathbf{z}^*(\mathbf{u})$ are performed conditional to the prior $\mathbf{z}(\mathbf{u}_i)$ (exploration holes) and posterior $\mathbf{z}(\mathbf{u}_\alpha)$ (the virtual infill-sampling). There are 480 (4 actual deposits x 4 sampling schemes x 30 cosimulations) simulated deposits to which the economics indicators would be applied. All those point-support cosimulations are then upscaled into 338 (26x13) blocks of dimension 15x15 metre. The block selection for each sampling scheme is based on the E-type mimicking the actual selection process for a mine operation. Finally, the profit per tonne mined and profit per tonne milled are calculated with expressions (5) and (6).

Considering the material stockpiled as waste, Figure 4 shows the histograms of P_{mined} for all sampling configurations for the first actual on two and three metre bench heights. First, all the schemes are profitable, i.e. no loss occurred, but some are more profitable than the others. It is also noticeable that the mean decreases with a sparser drilling pattern while the variance (and the coefficient of variation) increases. The 12x12 scheme is the most advantageous when considering both the efficiency and the uncertainty.

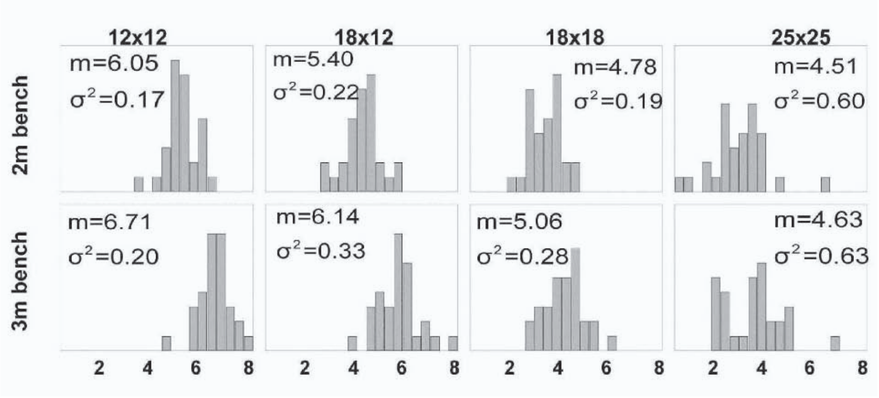


Figure 4 Histogram of profit per tones mined.

5.3 DISCOUNTING THE STOCKPILE AND EFFECTS ON DRILLING

In reality, the stockpile has a value that is neither ore nor waste. Stockpiling an economic block may be seen as ‘money in the bank’ without interest, thus inducing an opportunity cost. The cost of misclassifying an ore block as stockpile increases according to the stockpile strategy, i.e. when that block will be mined. From the banking analogy, the increase in cost depends on a discount rate and the number of years the material is to be stockpiled. The revenue generated from the stockpile, say to be mined in Δt years, is expressed in today’s dollar (t =0)

$$F_{t=0} = (1+i)^{-\Delta t} \; F_{t=\Delta t} \tag{7}$$

where i is the discount rate. A higher discount rate, i.e. a higher uncertainty about the stockpile strategy, will increase the misclassification cost.

The performance of the four infill drilling schemes is revisited by considering uncertainty in the stockpiled strategy modelled by discount rates. The cost of stockpiling marginal ore is investigated with six different discount rates on a period of 10 years. The profit generated by the stockpile is transformed into dollars in ten years time with a specific discount rate. With a discount rate of zero, suggesting no cost of opportunity, the stockpile is considered as ore. A high discount rate indicates that the stockpile lost all its value, thus it is considered as waste. All other discount rates are intermediate scenarios between these two “end-point” cases.

The profit per tonne mined is expressed in Australian dollar increment based on the currently used 12m x 12m sampling grid. The median profit per tonne mined for the four sampling scheme applied on actual #1 is shown in Figure 5. The upper left graph considers the stockpile as waste and the lower right graph as ore. The profit per tonne milled is shown in the same format in Figure 6.

For an increase of 8 cents per tonne in drilling cost between 12x12m and 25x25m, the profit per tonne mined improves up to $2 (35%) at a high cut-off (stockpile as waste) and of $0.50 at a lower cut-off (stockpile as waste). The 12x12m scheme is more profitable, even at a low cut-off (stockpile as ore). The 12x18m scheme does not decrease the profit too much and could also be appropriate for the deposit. The results seem insensitive to the bench height. For actual #1 the 2m is better than the 3m, the inverse is observed for the actual #2 therefore mining with 3m benches is appropriate.

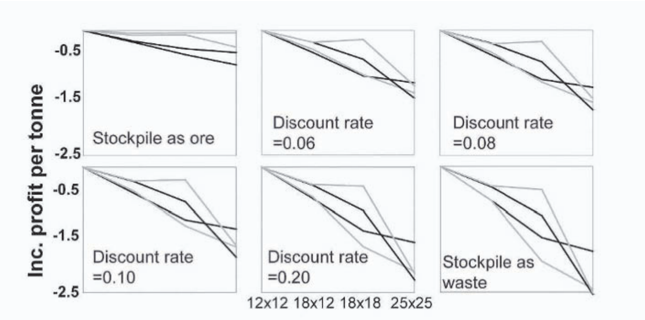


Figure 5 Median increment of profit per tonne mined. The median of the 12x12 metre scheme is set to zero. The black lines are for the actual deposits on two metre bench, the gray lines are for those on three metre bench.

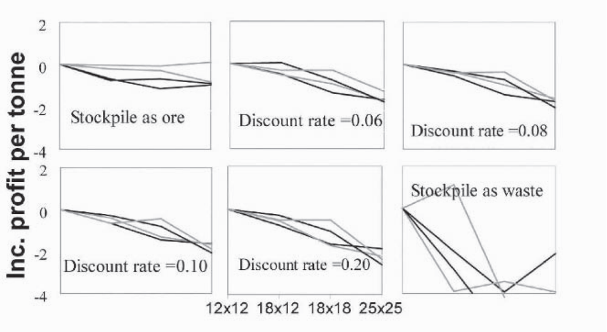


Figure 6 Median increment of profit per tonne milled. The median of the 12x12 metre scheme is set to zero. The black lines are for the actual deposit on two metre bench, the grey lines are for those on three metre bench.

6 Comments and Conclusions

This study shows that multivariate deposits can be efficiently simulated by first orthogonalising the attributes with the minimum/maximum autocorrelation factorisation. Once the simulated values are back-rotated, the cross-correlation at all lags between attributes is restored. The resulting realizations are then a better representation of the deposit and are therefore more appropriate for further processing.

The economic consequences of the drilling patterns on a multivariate deposit is then regarded on a large scale that takes into account some aspect of long-term planning, specifically with regards to the strategy and uncertainty related to the stockpiled material. The uncertainty of this material is translated into a discount rate, which indicates the risk of losing a profitable block when stockpiled. The performance of the drilling scheme can be assessed in a larger perspective than by the traditional misclassification parameters. This study demonstrates how the main mine planning decisions impact the lower level activities such as the spacing of the infill drilling.

Acknowledgment

The work presented herein was part of a research project funded by Anaconda Operations, Anglo Gold, BHP Billiton, Highlands Pacific, MIM Holdings (Xstrata), Pasminco, Rio Tinto and WMC Resources.

References

Almeida, A.S. and Journel, A.G. *Joint simulation of multiple-variables with a Markov-type coregionalization model*. Mathematical Geology vol 26, no. 5, 1994, p. 565-588.

Boucher, A. Conditional joint simulation of random fields on block support. Earth Sciences. Brisbane, University of Queensland: 2003

Chilès, J.-P. and Delfiner P., *Geostatistics modeling spatial uncertainty*., 1999

Desbarats, A. J. and Dimitrakopoulos, R. *Geostatistical simulation of regionalized pore-size distributions using min/max autocorrelation factors*. Mathematical Geology vol 32, no. 8, 2000 p. 919-942.

Dimitrakopoulos, R., 2003. Orebody *Uncertainty, Risk Assessment and Profitability in Recoverable Reserves, Ore Selection, and Mine Planning: Simulation Models, Concepts And Applications for the Mining Industry*. BRC, Brisbane, 342p.

Dimitrakopoulos, R. and Fonseca, M.B., *Assessing risk in grade-tonnage curves in a complex copper deposit, Northern Brazil, based on an efficient joint simulation of multiple correlated variables*. APCOM 2003*,* Cape Town

Dimitrakopoulos, R. and Luo, X., *Generalized sequential Gaussian simulation on group size v and screen effect approximations for large field simulations*. Mathematical Geology vol 36, no. 5, 2004, p. 567-591.

Goovaerts, P., *Geostatistics for Natural Resources Evaluation*, Oxford University Press, 1997.

Switzer, P. and Green A. A.., *Min/Max autocorellation factors for multivariate spatial imagery*. Stanford University, Department of Statistics. 1984.

Jaine, O J. *Mine planning at Murrin-Murrin - modelling to determine the optimum path.* in Proceedings of Twelfth International Symposium on Mine Planning and Equipment Selection (AusIMM: Melbourne), 2003

Ravenscroft, P.J*., Risk analysis for mine scheduling by conditional simulation*, Transaction of the institution of mining and metallurgy Section A, vol. 101, 1992, p. A104-A108

Verly, G. W., *Sequential Gaussian cosimulation; a simulation method integrating several types of information. In* Soares and Amilcar (Eds.) Geostatistics. Kluwer Academic Publishers, Dordrect, p. 543-554

THEORY OF THE CLOUD TRANSFORM FOR APPLICATIONS

ODD KOLBJØRNSEN and PETTER ABRAHAMSEN
Norwegian Computing Center, Oslo, Norway

Abstract. We present the multidimensional cloud transform and propose an estimator for the transform. The estimation procedure is based on scatter plot smoothing. The resulting transform does not introduce artificial discontinuities in the transformed data, which is a common problem for the traditional estimates. The method is compared to a traditional estimate in a synthetic example.

Key words: Non-Gaussian distribution, stochastic simulation, seismic conditioning

1 Introduction

Seismic data provide valuable information with high lateral resolution that improves reservoir models. Geophysical variables such as acoustic impedance, shear impedance and Poisson ratio are often available through out the reservoir as results of seismic inversions. The cloud transform, see Bashore et al. (1994), is a frequently used tool when incorporating one explanatory variable such as the acoustic impedance into the reservoir model. The multi dimensional cloud transform incorporates multiple explanatory variables in the transform. This is useful as elastic inversions that provide multiple geophysical variables now are quite common.

Traditional estimates of cloud transforms are constructed by introducing non-geological facies, e.g. impedance classes. This method introduces artificial discontinuities in the petrophysical simulations, and requires a large amount of well data in order to obtain a reliable result. When the explanatory data have multiple dimensions the traditional binning estimates will suffer due to lack of accuracy and precision of the estimates because the number of bins increases dramatically with the dimension.

In the current work we present the cloud transform using a probabilistic terminology and propose estimators for the cloud transform that is based on scatter plot smoothing. The major difference between the current approach and other scatter plot smoothing approaches, e.g. Xu and Journel (1995) and Deutsch (1996), is that we work with the cloud transform directly and do not consider the joint density. The resulting estimates yield continuous transforms. Asymptotic expressions for accuracy and precision are presented and discussed. Asymptotic convergence rates are obtained such that for a given target distribution the asymptotically ideal

O. Leuangthong and C. V. Deutsch (eds.), GeostatisticsBanff 2004, 45-54.

smoothing factor can be computed. The convergence rate of the estimator is the same as the convergence rate for the estimator of the density of the explanatory variables in the transform. Thus it converges faster than the kernel estimator of the joint density of response variable and explanatory variables.

A presentation of the cloud transform is given in section 2, the estimator and asymptotic properties are given in section 3, synthetic example with comparison of proposed estimators to the traditional estimate is given in section 4. At the end there is a discussion and concluding remarks in section 5 and 6 respectively.

In what follows the function f denotes a generic density, where the random variable(s) in question is implied by the argument(s) of f, e.g. $f(\boldsymbol{x})$ and $f(y)$ denotes the density of $\boldsymbol{X}$ and Y respectively. Bold letters are used to denote vectors, e.g. $\boldsymbol{x} \in \boldsymbol{\mathcal{R}}^d$. The function F denotes a cumulative distribution function the random variable in question is again implied by the argument, e.g. $F(y) = \text{Prob}(Y < y)$. Consequently $f(y|\boldsymbol{x})$ and $F(y|\boldsymbol{x})$ denotes the conditional pdf and cdf for Y given $\boldsymbol{X} = \boldsymbol{x}$ respectively. The quantile function that corresponds to the cdf F is denoted F^{-1} such that by definition $x = F^{-1}(F(x))$.

2 The cloud transform

The cloud transform is a conditional inverse probability transform. Let X and Y denote the explanatory and response variable respectively. Typically X is the acoustic impedance and Y is the porosity. A stochastic simulation from $f(x, y)$ can be obtained by the following algorithm:

Algorithm 1:

```
i)   Compute $F(x)$
ii)  Sample $u_1^* \sim$ Uniform [0, 1]
iii) Let $x^* = F^{-1}(u_1^*)$
iv)  Compute $F(y|x^*)$
v)   Sample $u_2^* \sim$ Uniform [0, 1] independent of $u_1^*$
vi)  Let $y^* = F^{-1}(u_2^*|x^*)$
vii) Return $(x^*, y^*)$.
```

The transform in step *iii)* is an inverse probability transform. The transform in step *vi)* is the cloud transform. The multi dimensional cloud transform denotes the case when the explanatory variable is multi dimensional, i.e. $y = F^{-1}(u|\boldsymbol{x})$. For example can the components of $\boldsymbol{x}$ be the acoustic and the Poisson ratio.

In a spatial setting the cloud transform is applied pointwise, i.e.

$$Y(\boldsymbol{s}) = F^{-1}(U(\boldsymbol{s})|\boldsymbol{x}(\boldsymbol{s})), \tag{1}$$

with $\boldsymbol{s}$ being the spatial reference, $Y(\boldsymbol{s})$ being the response field, $U(\boldsymbol{s})$ being a p-field and $\boldsymbol{x}(\boldsymbol{s})$ being the given explanatory field. A p-field has the property that the stationary distribution of a realisation of $U(\boldsymbol{s})$ is uniform on $[0, 1]$. A spatially correlated p-field can for example be obtained as $U(\boldsymbol{s}) = \Phi(Z(\boldsymbol{s}))$, with Φ being the standard normal cdf; and $Z(\boldsymbol{s})$ being a standard normal random field.

On a bounded domain the response field $Y(\boldsymbol{s})$ defined in expression (1) is almost surely continuous if $f(\boldsymbol{x}, y)$ is a density, $U^*(\boldsymbol{s})$ is almost surely continuous and $\boldsymbol{x}(\boldsymbol{s})$ is continuous almost everywhere.

The following algorithm use the cloud transform to reproduce the conditional distributions of $Y(\boldsymbol{s})$ given $\boldsymbol{x}(\boldsymbol{s})$: *Algorithm 2:*

```
i)   For all $\boldsymbol{s}$ in grid: compute $F(y|\boldsymbol{x}(\boldsymbol{s}))$
ii)  Sample a p-field $u^*(\boldsymbol{s})$ independent of $\boldsymbol{x}(\boldsymbol{s})$
iii) For all $\boldsymbol{s}$ in grid: let $y^*(\boldsymbol{s}) = F^{-1}(u^*(\boldsymbol{s})|\boldsymbol{x}(\boldsymbol{s}))$
iv)  Return $y^*(\boldsymbol{s})$.
```

In step *ii)* the term independent is used in terms of independent stationary distribution; i.e. all information regarding $Y(\boldsymbol{s})$ given by $\boldsymbol{x}(\boldsymbol{s})$ is given through the transform.

The cloud transform can also be used to reproduce joint multivariate distributions, by sampling in a sequential manner. The first variable is sampled according to an inverse probability transform; the next variables are sampled using the cloud transform given the previously sampled variables.

One can also imagine combinations of these two uses, by first simulating porosity given geophysical variables and next simulate permeability given geophysical variables and porosity.

The cloud transform become storage intensive as the dimension of the explanatory variable increase. Its hard store a cloud transform with a reasonable resolution if the dimension of the explanatory variable exceed four. In place of storing the transform one may consider to estimate it each time it is needed. The time requirement in this approach is prohibitive. In addition the number of data needed for a reliable estimate increase rapidly with the dimension.

3 Estimation of multi dimensional cloud transform

In an applied setting the cloud transform is unknown, and must be estimated from data. The estimator proposed here is based on the theory of kernel density estimation as presented in Silverman (1986), main results are summarised below. Other methods of density estimation see e.g. Donoho et al. (1996) and more refined approaches to kernel smoothing see e.g. Sain and Scott (1996), can also be developed into the setting of the cloud transform.

3.1 DENSITY ESTIMATION

Let $\boldsymbol{X}_1, \boldsymbol{X}_2, ..., \boldsymbol{X}_n$ be a given multivariate data set whose underlying density is to be estimated. The kernel density estimator of the joint density is then

$$\hat{f}(\boldsymbol{x}) = \frac{1}{nh^d} \sum_{i=1}^{n} k_d \left(\frac{\boldsymbol{x} - \boldsymbol{X}_i}{h} \right), \tag{2}$$

with h being the bandwidth; and the kernel $k_d : \boldsymbol{\mathcal{R}}^d \rightarrow \boldsymbol{\mathcal{R}}$ being a radially symmetric unimodal probability density function such that

$$\int_{\boldsymbol{\mathcal{R}}^d} x_i x_j k_d(\boldsymbol{x}) d\boldsymbol{x} = \delta_{ij},$$

where δ_{ij} is one if $i = j$; zero otherwise. Define further the constant

$$\beta_d = \int_{\boldsymbol{\mathcal{R}}^d} |k_d(\boldsymbol{x})|^2 d\boldsymbol{x},$$

that is specific for the kernel k_d.

A standard argument using Taylor expansions yields the asymptotic expression for bias,

$$\mathrm{E}\left\{\hat{f}(\boldsymbol{x})\right\} - f(\boldsymbol{x}) \doteq \frac{h^2}{2} \nabla^2 f(\boldsymbol{x}), \tag{3}$$

with ∇^2 being the Laplace operator in $\boldsymbol{\mathcal{R}}^d$. The asymptotic variance is

$$\mathrm{Var}\left\{\hat{f}(\boldsymbol{x})\right\} \doteq \frac{\beta_d}{nh^d} f(\boldsymbol{x}). \tag{4}$$

Combining the two yields the mean squared error

$$\begin{aligned} \mathrm{MSE}\left\{\hat{f}(\boldsymbol{x})\right\} &= \left[\mathrm{E}\left\{\hat{f}(\boldsymbol{x})\right\} - f(\boldsymbol{x})\right]^2 + \mathrm{Var}\left\{\hat{f}(\boldsymbol{x})\right\} \\ & \quad \frac{h^4}{4}|\nabla^2 f(\boldsymbol{x})|^2 + \frac{\beta_d}{nh^d} f(\boldsymbol{x}). \end{aligned} \tag{5}$$

From this expression one obtain the optimal rate of convergence for the bandwidth being,

$$h_{\text{opt}} \sim n^{-1/(d+4)} \tag{6}$$

yielding the convergence rate of the mean squared error,

$$\mathrm{MSE}\left\{\hat{f}(\boldsymbol{x})\right\} \sim n^{-4/(d+4)}. \tag{7}$$

The integrated mean square error (IMSE) is a common measure of error in density estimation and is used to identify a common bandwidth for all $\boldsymbol{x} \in R^d$. It is however not possible to find a universal bandwidth that is applicable of all densities since the MSE and IMSE depend on the target density, see expression (5).

Consider also estimation of the cumulative distribution in one dimension, $F(x)$, using the ordinary count estimator,

$$\hat{F}(x) = \frac{\sum_{i=1}^n I(X_i \leq x)}{n}, \tag{8}$$

with $I(X_i \leq x)$ being one if its argument is true zero otherwise. This estimator is unbiased and has variance according to the estimator of the probability in a binomial distribution. The mean squared error is thus identical to the variance which is

$$\mathrm{Var}\left\{\hat{F}(x)\right\} = \frac{F(x)\,[1-F(x)]}{n}. \tag{9}$$

The convergence rate for the MSE of the count estimator is hence of order n^{-1}. This should be compared with the convergence rate of the density estimator, see expression (7). In one dimension the convergence rate for the density is $n^{-1+1/5}$. The convergence rate of the cdf corresponds to $d = 0$ in expression (7).

3.2 CLOUD TRANSFORM ESTIMATION

Let $(Y_1, \boldsymbol{X}_1)$, $(Y_2, \boldsymbol{X}_2)$, ..., $(Y_n, \boldsymbol{X}_n)$ be a multivariate dataset for which the cloud transform is estimated with Y and $\boldsymbol{X}$ being the response and explanatory variable respectively. The kernel estimator of the joint density is then,

$$\hat{f}(\boldsymbol{x}, y) = \frac{1}{n\,h^d h_y} \sum_{i=1}^{n} k_d\left(\frac{\boldsymbol{x}-\boldsymbol{X}_i}{h}\right) \cdot k_1\left(\frac{y-Y_i}{h_y}\right), \tag{10}$$

where the kernel is separated for $\boldsymbol{x}$ and y; and h_y is the bandwidth used for the response variable. The target for the estimation is the conditional cumulative distribution $F(y|\boldsymbol{x})$. When using the density estimate in expression (10) one can obtain the estimator of $F(y|\boldsymbol{x})$ as,

$$\hat{F}(y|\boldsymbol{x}) = \frac{\sum_{i=1}^{n} k_d\left(\frac{\boldsymbol{x}-\boldsymbol{X}_i}{h}\right) \cdot K_1\left(\frac{y-Y_i}{h_y}\right)}{\sum_{i=1}^{n} k_d\left(\frac{\boldsymbol{x}-\boldsymbol{X}_i}{h}\right)}, \tag{11}$$

with $K_1(y) = \int_{-\infty}^{y} k_1(t)\,dt$. The bias in the estimator in expression (11) has the complexity

$$\mathrm{E}\{\hat{F}(y|\boldsymbol{x})\} - F(y|\boldsymbol{x}) \;\sim o(h^2 + h_y^2)\,.$$

The asymptotic variance has the complexity

$$\mathrm{Var}\left\{\hat{F}(y|\boldsymbol{x})\right\} \;\sim o\left(\frac{1}{nh^d}\right).$$

The bound for the asymptotic variance is independent of h_y. This is intuitively explained by the fact that $K_1(y/h_y)$ in expression (11) is bounded whereas $k_1(y/h_y)/h_y$ in expression (10) is unbounded when h_y approaches zero. The usual trade off between bias and variance is not needed in the direction of the response variable. Thus let $h_y = 0$ an introduce the unnormalised conditional cdf of Y given $\boldsymbol{X}$,

$$G(y;\boldsymbol{x}) = \int_{-\infty}^{y} f(t,\boldsymbol{x})dt = F(y|\boldsymbol{x})f(\boldsymbol{x}).$$

The asymptotic bias for the case of $h_y = 0$ is

$$\mathrm{E}\{\hat{F}(y|\boldsymbol{x})\} - F(y|\boldsymbol{x}) \doteq \frac{h^2}{2f(\boldsymbol{x})}\left[\nabla_{\boldsymbol{x}}^2 G(y;\boldsymbol{x}) - F(y|\boldsymbol{x})\nabla_{\boldsymbol{x}}^2 f(\boldsymbol{x})\right], \tag{12}$$

and the asymptotic expression for the variance is

$$\operatorname{Var}\left\{\hat{F}(y|\boldsymbol{x})\right\} \doteq \frac{\beta_d}{nh^d\, f(\boldsymbol{x})}\, \left[F(y|\boldsymbol{x})\,(1 - F(y|\boldsymbol{x}))\right]. \tag{13}$$

It is interesting to compare this variance with the one obtained for estimating empirical cumulative distributions in 1D. The factor $[F(y|\boldsymbol{x})\,(1 - F(y|\boldsymbol{x}))]\,/[nh^d f(\boldsymbol{x})]$ can be interpreted as the binomial uncertainty given $[nh^d f(\boldsymbol{x})]$ data, see expression (9). The factor β_d is related to the kernel smoothing, see expression (4).

By combining the bias in expression (12) and the variance in expression (13) to the mean squared error one see that the optimal rate of convergence for the bandwidth is obtained by

$$h_{\text{opt}} \sim n^{-1/(d+4)},$$

yielding the convergence rate for the mean squared error to be

$$\text{MSE}\{\hat{F}(y|\boldsymbol{x})\} \sim n^{-4/(d+4)}. \tag{14}$$

This is the same rate of convergence as obtained for density estimation, see expression (7), but in expression (14) the dimension d refers to the dimension of the explanatory variables.

Note in particular that an estimator of the cloud transform that is based on the optimal kernel density estimator will have the convergence rate $n^{1/(d+5)}$, which is suboptimal.

The factor $1/f(\boldsymbol{x})$ which occur both in expression (12) and (13) is large in the flanks of $f(\boldsymbol{x})$. This factor may be reduced by transforming the explanatory variable to be approximately uniform on $[0,1]^d$. In the case of a one dimensional explanatory variable, the rank transform is uniform. In higher dimensions it is possible to obtain approximate uniform distributions by sequentially estimating the conditional transforms in the same manner as for the cloud transform, but applying them to the explanatory variables. This transform reduce the variance in the estimate but unfortunately the bias is increased trough the factor $\nabla^2_{\boldsymbol{x}} G(y;\boldsymbol{x})$. The advantage is that the transform remains stable at the flanks. When the cloud transform is used to model spatial phenomena the histogram based on well logs have a smaller support than the histogram of the full field, due to the number of samples. It is therefore importance to have a reliable estimate of the cloud transform also towards the flanks of the distribution of the explanatory variable.

Note that the kernel estimator of the cloud transform corresponds to a density estimator for the explanatory variables. If the stationary distribution of the explanatory variable is an exhaustive sampling of this distribution, the marginal histogram of the dependent variables is exactly reproduced in the full field.

3.3 SPECIAL CASES

It is interesting to investigate the proposed estimator for cases in which one can see the effect directly.

In the trivial case where there is no explanatory variable, the estimator is identical to the empirical cdf of the response variable, see expression (8).

If the response variable is independent of the explanatory variable, the estimator introduces local bias for the simulated response variable. However the stationary distribution of the response variable will still be reproduced. In this case it is obvious that an infinite bandwidth is optimal for the explanatory variable.

When the response variable is discrete, the estimator is identical to estimates obtained by kernel density estimation for each level of the response variable. The kernel density estimates for all classes have a common bandwidth. The probability of a class at a given value of the explanatory variable is proportional to the density estimate weighted with the number of data in this class.

If there is a functional relationship between the explanatory and the response variables, the estimator blurs this relation. This introduces artificial uncertainty in the predictions. The obvious choice in this case is to estimate this deterministic relation instead of introducing the cloud transform which is a stochastic transform.

4 Example

The properties of the estimators are investigated in a synthetic example where a relation between acoustic impedance and porosity is considered. In Figure 1 the scatter plot of the data that are used to estimate the transform is displayed together with the cloud transform. A vertical line in the cloud transform yields a cumulative distribution for the porosity increasing monotonically from zero at low porosity values to one for high porosity values. In the figure both extreme ends are coloured white in order to highlight the active region.

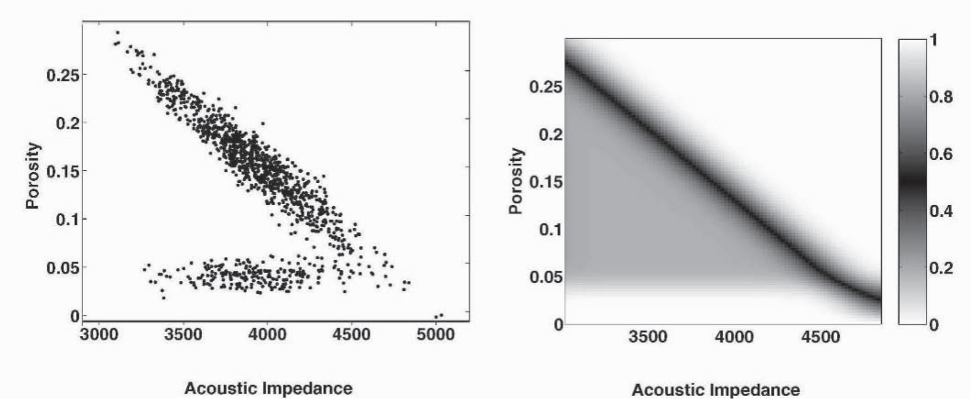


Figure 1. Data and original transform. On the left is the scatter plot of the 1200 well observations that are used to estimate the transform. On the right is the original cloud transform of the joint distribution of acoustic impedance and porosity. All cumulative distributions are zero for low porosity values and one for high values.

The binned estimator is compared with two estimators based on scatter plot smoothing. The first use the basic variable, i.e. acoustic impedance, the second use the rank transform of the basic variable as explanatory variable in the cloud transform. In Figure 2 the three estimated cloud transforms are displayed together with the original transform. Both estimates based on scatter plot smoothing are continuous whereas the binned estimate has clear discontinuities as the acoustic impedance crosses the boarder between bins. The binned estimate and the untransformed scatter plot smoother become unstable at the ends. This is a result of the high variance in the estimate in the extreme ends. The transformed scatter plot smoother is stable at the ends, but the bias is evident in the figure.

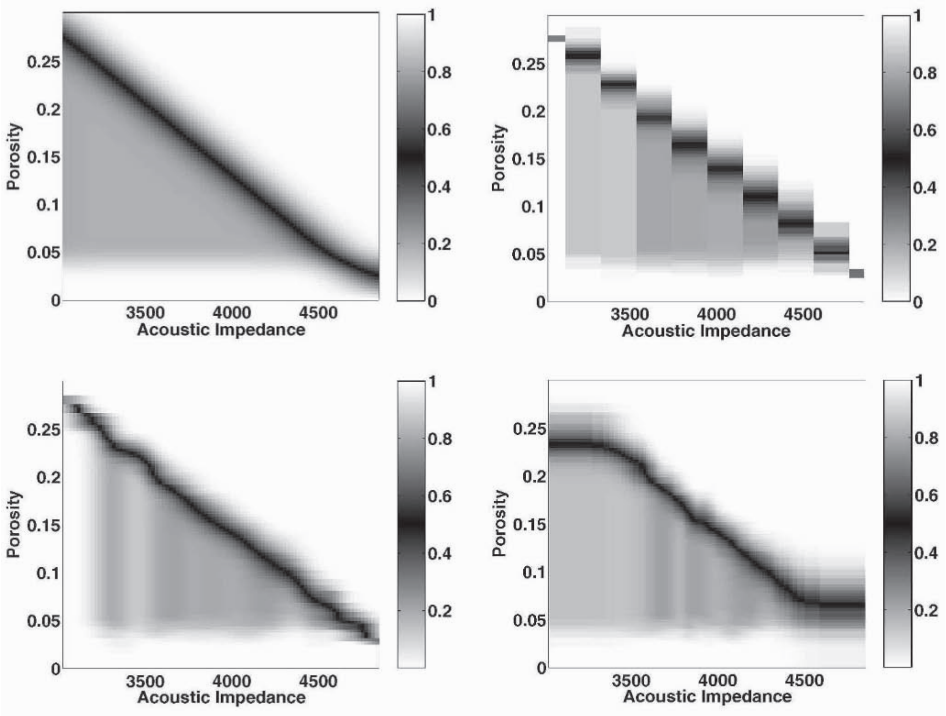


Figure 2. Original and estimated transforms. On the top left is the original cloud transform i.e. the target of estimation, top right is the binned transform, on bottom left is the estimate based on scatter plot smoothing, bottom right is the estimate based on a rank transform of the acoustic impedance. All three estimates are based on the 1200 well observations displayed in the scatter plot in Figure 1.

In order to compare statistical properties of the estimators of the cloud transform the root integrated mean square error,

$$\text{RIMSE}\left\{\hat{F}(y|x)\right\} = \left(\int_{\mathcal{R}} \text{MSE}\{F(y|x)\}\, dy\right)^{1/2},$$

is computed for each value of the acoustic impedance. The mean squared error is approximated by Monte Carlo integration using the following procedure. Generate 1000 independent data sets all consisting of 1200 data pairs. For each data set estimate the transform and compute the squared deviation between this and the true transform. The average of the 1000 squared deviations is the approximation to the mean squared error. The results for the three estimators are displayed in Figure 3, the density of the acoustic impedance, i.e. $f(x)$, is overlaid in the figure.

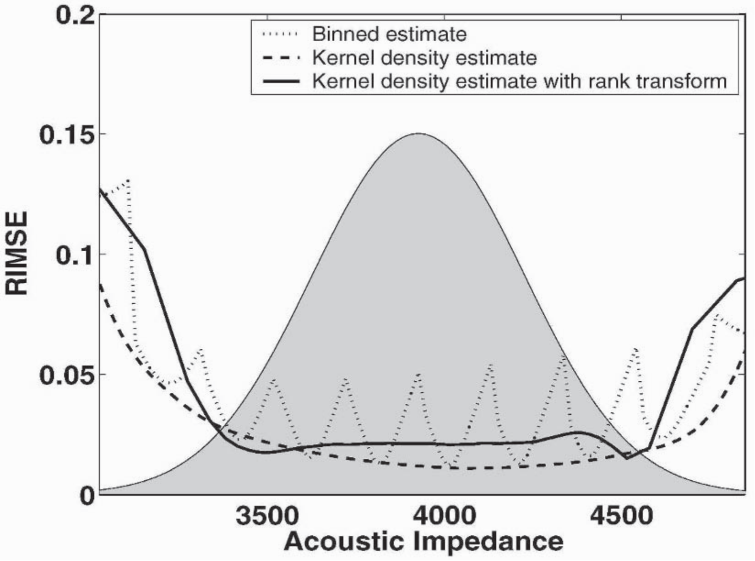


Figure 3. Pointwise root integrated mean square error. The root integrated mean square error for three estimators considered. Overlaid is the density of acoustic impedance. The range in the figure is about six times the standard deviation for acoustic impedance.

In terms of root integrated mean square error both estimates based on scatter plot smoothing outperform the binned estimate. The scatter plot smoother that use the basic variable is better than the one based on the rank transform.

5 Discussion

The cloud transform can be used to reproduce any multivariate distribution; however the spatial dependence is hidden in the p-field. A scatter plot such as the one in Figure 1 may indicate an underlying dichotomous random field. It is not obvious how to create a p-field with desired spatial properties, e.g. channels. However if the wells are dense compared with the correlation length of the random fields a p-field originating from a transformed Gaussian field may be satisfactory. An alternative approach is to build a facies model with the desired spatial properties and build separate cloud transforms for petrophysical modelling within each facies.

Scatter plots that are used for estimation of the cloud transform usually come from well observations. This will result in data that are correlated and not independent which is assumed in the calculations above. This will most likely have the effect that the variance of the estimator is larger than given in expression (13) above. The scatter plot may also come from rock physics simulations. In which case it is likely that the data are independent and the results are strictly valid.

6 Concluding remarks

We have proposed two estimators for the cloud transform. Both estimators are based on scatter plot smoothing and result in continuous estimates. The optimal bandwidth of the estimator has the same convergence rate as for the density estimation in the space of explanatory variables. There is however no need to smooth in the direction of the response variable as this introduces additional bias.

In a test example both proposed estimators are found to perform better than the traditional binning estimate in terms of root integrated mean squared error, also the estimators are more appealing visually as they preserve continuity in the estimate. The estimator based on the basic explanatory variable is the best of the two estimators in terms of root integrated mean squared error. This estimator does however have large variance at the ends of the interval resulting in an unstable estimate. This can be unfortunate, see discussion in end of section 3.2.

When choosing estimator for the cloud transform one should not only consider its theoretical properties, but also how the resulting estimate will be applied. The authors prefer a slightly higher bias in order to preserve good properties, i.e. smoothness, of the estimated transform at the flanks of the distribution of the explanatory variable.

Acknowledgements

The authors acknowledge Alistair MacDonald of Roxar ASA. The Norwegian research council has in part provided the funding for this work.

References

Bashore, W M.; Araktingi U.G.; Levy M; Schweller U.G. *Importance of a Geological Framework for Reservoir modelling and subsequent Fluid -Flow Predictions* AAPG Computer application in geology., No.3; Jeffery M. Yarus and Richard L Chambers eds.; 1994.

Deutsch C.V. *Constrained Smoothing of Histogram and Scatter plots with Simulated Annealing.* Technometrics, Vol. 38, No. 3,pp. 266-274,1996.

Donoho D.L.; Johnstone I.M.; Kerkyacharian G. and Picard D. *Density Estimation by Wavelet Thresholding* The Annals of Statistics, Vol. 24, No. 2, pp. 508-539, 1996.

Sain S. R.; Scott D. W. *On Locally Adaptive Density Estimation* Journal of the American Statistical Association, Vol. 91, No. 436, pp. 1525-1534, 1996.

Silverman B.W. *Density Estimation for Statistics and Data Analysis.*Chapman & Hall 1986.

Xu, W. and Journel, A.G. *Histogram and Scattergram Smoothing Using Convex Quadratic Programming* Mathematical Geology.Vol 27 , No.1; pp. 83-103 ; 1995.

PROBABILITY FIELD SIMULATION: A RETROSPECTIVE

R. MOHAN SRIVASTAVA[1] and ROLAND FROIDEVAUX[2]
[1]*FSS Canada Consultants, Toronto, Canada*
[2]*FSS Consultants SA, Geneva, Switzerland*

Abstract.

The practical advantages and theoretical disadvantages of P-field simulation are reviewed in the light of more than a decade of application and research since it was first introduced. A case study example highlights the enduring attractions of the algorithm: its flexibility and speed.

1 Introduction

When first introduced, probability field simulation was well-suited to certain types of problems that were not well handled by other simulation algorithms available. In particular, it adapted well to the situation where *a priori* local distributions were available. As it rapidly gained practical acceptance, largely because of its speed, "P-field" simulation was also dismissed by some as a procedure lacking a proper theoretical foundation — more of a clever algorithmic trick than a properly conceived approach to stochastic spatial simulation.

In the past decade, the advantages and shortcomings of the procedure have been illuminated through continued widespread application and theoretical research. This paper begins with an overview of the theoretical background and the usual practical implementation of P-field simulation. It then discusses theoretical concerns and assesses their practical implications. A mining case study example illustrates two enduring strengths of P-field simulation: flexibility and speed.

2 Overview and implementation

Let $F[\mathbf{u}; z]$ denote the cumulative distribution function (cdf) at location $\mathbf{u}$ of an attribute Z. Any simulated value, z_{sim}, represents a specific quantile of this local cdf: the z-value at which $F[\mathbf{u}; z]$ reaches a probability $p(\mathbf{u})$:

$$z_{\mathrm{sim}} = F^{-1}[\mathbf{u}; p(\mathbf{u})] \tag{1}$$

The p values are not spatially independent; this would preclude reproduction of almost any desired spatial autocorrelation in Z. Instead, the p values must be regarded as a realization of a random function $P(\mathbf{u})$, and simulated with an appropriate pattern of spatial continuity.

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 55-64.

P-field simulation therefore proceeds as follows:

1. Generate a non-conditional realization of $P(\mathbf{u})$, i.e. a grid of spatially autocorrelated values that are uniformly distributed between 0 and 1.[1]
2. Use $P(\mathbf{u})$ to sample the local cdf $F[\mathbf{u}; z]$.

This procedure ensures that two of the common goals of conditional simulation are met: conditioning data are honored, as is the target global distribution. The global distribution of $Z(\mathbf{u})$ is honored because the local cdfs are sampled using $U[0,1]$ values. As long as local cdfs correctly model local distributions of uncertainty, sampling these with $U[0,1]$ values will preserve the global distribution. Conditioning data are honored because local cdfs collapse to a spike at data locations. Regardless of the probability value used to sample these zero-width distributions, the simulated value will match the conditioning data value.

The third common goal of conditional simulation, the reproduction of the variogram of $Z(\mathbf{u})$, is not exactly guaranteed. Since the $p(\mathbf{u})$ values are spatially autocorrelated, the $z_{\text{sim}}(\mathbf{u})$ will also be spatially autocorrelated, but the precise nature of the autocorrelation of the z_{sim} values is not directly controlled. The resulting variogram of the z_{sim} values will not necessarily reflect the intended target Z variogram model. As discussed later, it will often be very close to the desired target but there are situations in which, despite having *some* spatial continuity, the z_{sim} values do not have exactly the desired pattern of spatial continuity.

3 Theoretical considerations

The first P-field papers (Srivastava, 1992; Froidevaux 1993) focused on algorithmic details; little theoretical justification was provided and the acceptance of the procedure was due to its practical success. Theoretical investigations soon followed, however, and links between the P-field approach and other conditional simulation methods were eventually elucidated (e.g. Journel and Ying, 2001).

Although Journel (1995) proved that, in the absence of conditioning data, P-field simulation correctly reproduces univariate and bivariate properties of Z, Pyrcz and Deutsch (2001) pointed out that: i) if a stationary covariance model is used for P, the covariance of Z is not stationary and is biased in the vicinity of conditioning data and, ii) conditioning data usually appear as local extremes in the realizations.

3.1 INFERRING THE LOCAL CDFS

P-field simulation does not concern itself with the determination of the local cdfs; it considers them to have already been established. The origin of the local cdfs

[1] This is usually done by generating non-conditional Gaussian values, $Y(\mathbf{u})$, and then using the inverse of the cumulative Gaussian distribution to transform the Y values to P values: $P(\mathbf{u}) = G^{-1}[Y(\mathbf{u})]$. The generation of spatially autocorrelated Gaussian values can be done extremely rapidly using fast Fourier transform (FFT) algorithms or by efficient moving averages. In applications where the local cdfs are Gaussian, the transformation from Y values to P values can be skipped and the Y values are simply linearly transformed to a Gaussian distribution with the proper mean and variance.

does play a role in the theoretical analysis of the spatial structure of P. It is useful, therefore, to elucidate some common cases for the determination of local cdfs and to discuss how these impinge on the variogram model for P:

Case 1: Local cdfs not locally data-conditioned but are identified instead with the prior marginal distribution of Z: $\text{Prob}\{Z(\mathbf{u}) < z\} = F(z)$.

Case 2: Local cdfs are not locally data-conditioned but are identified instead with non-stationary prior distributions of Z: $\text{Prob}\{Z(\mathbf{u}) < z\} = F(\mathbf{u}; z)$.

Case 3: Local cdfs are estimated from existing sample data using an appropriate geostatistical technique: $\text{Prob}\{Z(\mathbf{u}) < z\} = F[\mathbf{u}; z|(n)]$.

The single most important issue here is conditioning to sample data. This will have a direct impact on the inference of the variogram model.

3.2 P-FIELD VARIOGRAM MODEL: STATIONARY OR NOT?

P-field simulation usually uses a stationary variogram model for P. This normal practice follows from the original suggestion of Froidevaux (1993): that the P variogram be modelled from the experimental variogram of the uniform transform of the available data. As Pyrcz and Deutsch (2001) pointed out, howeer, if the Z values are assumed to be second-order stationary, then the use of a stationary variogram model for P is inconsistent. If P is defined as

$$P(\mathbf{u}) = F[\mathbf{u}; Z(\mathbf{u})] \tag{2}$$

using data-conditioned local cdfs, then second-order stationarity of Z entails lack of second-order stationarity for P. Cassiraga (1999) has shown that the range of autocorrelation of P is linked to the spatial density of the conditioning data.

To date, theoretical analysis of P-field simulation has proceeded from the assumption that the Z values have second-order stationarity, and that the P values are defined using Equation 2 above. One could, however, take a different approach: assume that the P values have second-order stationarity and that the Z values are defined using Equation 1. P and Z play complementary roles in Equations 1 and 2, and the results of Cassiraga (1999) can be extended to demonstrate that if we choose a random function model in which the P values are second-order stationary, then the consequence is that the Z values cannot be; or, as also pointed out by Pyrcz and Deutsch (2001), the stationarity of the P-field covariance makes the covariance structure of Z dependent on the nearby conditioning data.

So with two alternate random function models — one that is better researched and that chooses second-order stationarity for Z; the other that chooses second-order stationarity for P and whose theory has barely been explored — the question arises: which one is more appropriate? Though the tradition of geostatistics has been to choose second-order stationarity as a Z property, it is worth considering the pro's and con's of bestowing this property on P instead.

It is clear from the construction of the P-field that the P values are, globally, first-order stationary; if they are not, then the local cdfs do not properly quantify the local probability distribution of Z. It is equally clear, from practice, that in most interesting earth science applications, first-order stationarity of Z is a questionable choice. Geostatistics has adopted the good practice of using local search neighborhoods so that the dependence on stationarity becomes local; but the practical success of local customization of estimation and simulation parameters is consistent with the view that an assumption of global first-order stationarity is rarely appropriate for Z.

Moving from the consideration of first-order stationarity to second-order stationarity, if the Z values are not first-order stationary, why does it make sense to assume that they are, globally, second-order stationary? Might it not be better to assign the property of second-order stationarity to a random variable, P, that is known, by construction, to be globally first-order stationary?

The technical literature on P-field simulation has elucidated the fact that P and Z cannot both be second-order stationary. Research remains to be done on the theoretical consequences of the user's choice on which of the two complementary random variables this property will be assigned to.

3.3 LOCAL EXTREMES AT DATA LOCATIONS

When hard data are used to locally condition cdfs, a sample at $\mathbf{u}$ will typically have a very strong influence on the cdf at an adjacent location, $\mathbf{u}'$. If the local cdf $F[\mathbf{u}'; z|(n)]$ has been estimated geostatistically, then its mean will tend to be very close to the adjacent data value, $z(\mathbf{u})$, and its variance will be small. Given this situation, if the p values in the vicinity of $\mathbf{u}$ are significantly less than 0.5, then $z(\mathbf{u})$ will be a local maximum in the realization. Conversely, if the nearby p values are larger than 0.5, then $z(\mathbf{u})$ will be a local minimum. The conditioning data, $z(\mathbf{u})$, will not be noticeable as a locally extreme artifact in the realization only if the nearby probability field values are around 0.5.

4 Discussion

4.1 DECOUPLING CDF ESTIMATION FROM SAMPLING

The practical advantage of P-field simulation stems from the decoupling of the sampling of cdfs from their estimation. As with other geostatistical simulation procedures, the local cdfs in a P-field approach can be established through some form of kriging; they can also be derived directly from secondary information. In many petroleum applications, for example, geophysics or petrophysics can provide constraints on rock properties such as porosity and permeability, and on structural properties such as thickness and depth to top of reservoir. In such situations, local cdfs can be built directly from geophysical data and no kriging is required; all that remains is the appropriate sampling of the geophysically-derived local cdfs.

In studies that involve resource estimation, the decoupling of cdf estimation from cdf sampling has another benefit: it is easier to ensure that simulated outcomes do not imply outlandish or aberrant resource estimates. Though the concept that simulations fluctuate around the expected value is well understood theoretically, in practice it can be hard to ensure that the average of many realizations is suitably close to an already-calculated resource estimate. There are many situations, especially in mining applications, where conditional simulations are being considered (for grade control, for example, or for blending studies) and where a well accepted and carefully developed resource block model already exists. The use of a P-field approach that incorporates previously accepted and trusted local cdfs avoids the embarassment and confusion that results when simulated outcomes depart, significantly on average, from the “best estimate” of the global resource.

4.2 HONORING THE VARIOGRAMS

As noted above, the definition and inference of the P variogram is theoretically troublesome if the Z values are assumed to be second-order stationary. The practical impact of this issue is, however, usually minor. With the real goal being reproduction of the Z variogram; the P variogram is an intermediate steppingstone. Even if the P variogram is theoretically ill-defined, the user can still adopt a variogram model based on analysis of the uniform transform of Z and can adjust this model if the resulting variogram of the z_{sim} values is unacceptable.

Luster (1985) discussed departures between target variogram models and experimental variograms of realizations. He noted that, by virtue of being conditioned by hard data, realizations have a pattern of spatial continuity whose mid- and long-range structure is controlled not by the variogram model but rather by available data. In practice, the critical aspect of the P variogram model is, therefore, its behavior at short distances (up to the nominal spacing of data). With the short-scale characteristics of the P variogram model well chosen, especially directional anisotropy and relative nugget effect, the results of P-field simulation are usually well within the fluctuations normally tolerated in conditional simulation studies.

Compared to sequential methods, P-field simulation is more successful at creating realizations with very low nugget effects and strong short-scale continuity (such as those typical of thickness or top of structure in petroleum applications). The realizations from sequential methods often have too much short-scale variability[2] and need to be post-processed to remove such artifacts (e.g. Tran, 1994).

4.3 LOCAL EXTREMES AT DATA LOCATIONS

To solve the problem of local extremes at data locations, Goovaerts (2002) proposed the use of a conditional probability field with fixed probability values of 0.5 at data locations. This entails a preferential sampling of the central part of the cdfs in the immediate vicinity of conditioning data. Although this method removes the artifacts, it does so at the expense of execution speed, which is one of the most attractive features of P-field simulation. Moreover, the justification and the consequences of forcing an arbitrary fixed p value still remain to be explored.

[2] A consequence of unstable kriging weights caused by strong screen effects in the end-stages of sequential simulation on a dense regular grid.

A practical application for which local extremes at data locations are clearly undesirable is flow and transport modelling. If wells or bore holes coincide with local minima and maxima in the permeability field, attempts to predict flow and transport may be seriously biased. In this sense, the P-field artifact of local extremes is similar to the "striping" or "banding" artifact often seen in realizations from the turning bands method and in realizations from sequential methods that do not randomize the sequential path. While such artifacts may not have any practical impact in certain types of studies, they may be serious flaws in others.

5 Case study: uncertainty on a mineralized envelope

Studies of mineral resource estimates typically incorporate a "mineralized envelope", an outer bounding limit beyond which grades are not estimated. Many case studies have demonstrated that the failure to adequately constrain the domain within which grades are estimated can lead to very unrealistic block models that overstate the tonnage of mineralized material, with peripheral grades being overestimated and grades in the heart of the deposit being underestimated.

Though some kind of mineralized envelope is necessary, the traditional approach, unfortunately, is to treat this boundary as deterministic. The limits of mineralization identified in drill holes typically serve as control points for a 3D solid or "wireframe". With the mineralized envelope thus frozen, the impact of the uncertainty of this envelope on resource estimates is very difficult to quantify. Even when simulation is used to study grade fluctuations within the envelope, the additional uncertainty due to the wireframe definition itself is rarely addressed.

Figure 1 shows an example of a simple wireframe constructed from exploration holes drilled on a 100m grid. These holes identify a deposit that lies in a shear zone between two faults, with a sharp hangingwall contact that can usually easily be correlated from hole to hole across the deposit. The footwall contact, which is more diffuse, does not appear to be a structural contact and is not clearly associated with any particular geological characteristic. In places, the footwall coincides with feldspathic alteration; in other places, it occurs where the intensity of shearing drops.

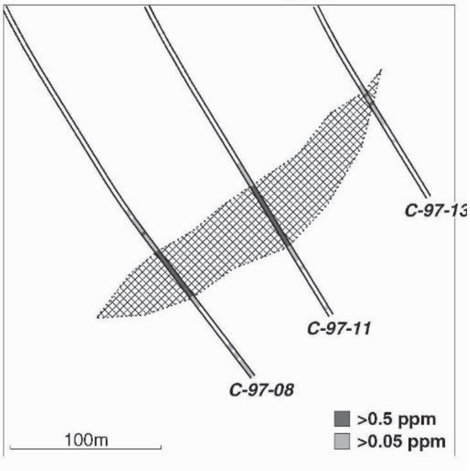


Figure 1. Drill holes and interpretation of mineralized envelope from initial drilling.

All three holes on the north-facing section shown in Figure 1 intersect anomalous mineralization which, for this project, was defined as more than five consecutive meters of drill core with total precious metals (TPM) in excess of 0.5 g/t. Small changes in the grade threshold or the length of interval have little impact on the definition of the hangingwall contact but have a more appreciable impact on the footwall.

The wireframe developed from initial drilling is necessarily simplistic, little more than a schematic cartoon that approximates the deposit's heart. In a second drilling campaign, in-fill holes were drilled from a development drift to penetrate the deposit from the west. Figure 2 shows the new holes with their mineralized intercepts, along with the old drill hole data and the mineralized envelope from Figure 1. The original wireframe provided good predictions of down-hole depth to the hangingwall, but its predictions of depth to the footwall are less precise.

Figure 3 shows the interpreted mineralized envelope, updated to honor all data currently available. With more closely spaced data, the shape of the wireframe has become slightly more complex. Though this new interpretation is an improvement, it is still far from perfect. If even more closely spaced holes were available, new short-scale complexities would be discovered in the shape of the mineralized zone. Rather than using the outline in Figure 3 as a single deterministic boundary for purposes of resource estimation, we would like to run several resource estimates, each one with a different but plausible version of the mineralized envelope, to study the impact on resources of uncertainty on the shape of the deposit. In conjunction with conditional simulations of TPM grades within the mineralized envelope, this will help determine whether or not additional definition drilling is required.

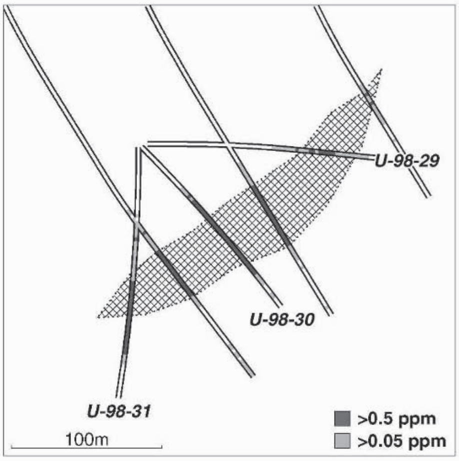


Figure 2. Drill holes after underground development drilling, with old interpretation of mineralized envelope.

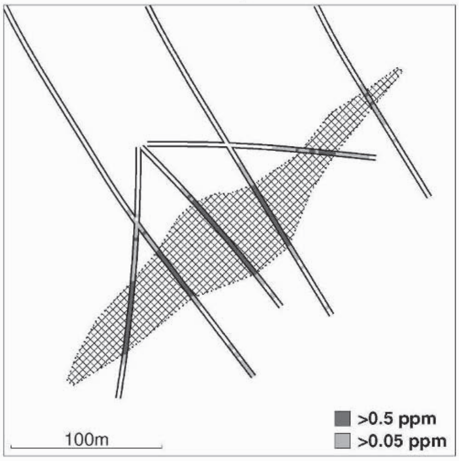


Figure 3. Drill holes after underground development drilling, with new interpretation of mineralized envelope.

P-field simulation has been used to produce alternate versions of the mineralized envelope, each one of which honors the drill hole data from the first two years. The attribute being simulated is $\Delta D(\mathbf{u})$, the deviation[3] of the true (but largely unknown) mineralized envelope from the current working interpretation shown in Figure 3. Close to existing drill holes, the current working interpretation is reliable and $\Delta D(\mathbf{u})$ is close to 0. As we move farther away from existing holes, the deviations between the actual and predicted surfaces will tend to become larger.

[3] Measured orthogonal to the wireframed surface, with the sign determining whether the deviation is outward (+) or inward (–).

For the 62 holes drilled in the second year, and which intersected the mineralized zone, Figure 4a shows the differences between actual depth to the hangingwall and predicted depths as a function of distance from a hole drilled in the first year. This plot, which shows us actual historical values of $\Delta D(\mathbf{u})$, can be used to calibrate possible future fluctuations. The dashed line in Figure 4a shows a model of ± one standard deviation of $\Delta D(\mathbf{u})$ as a function of distance from an existing drill hole; the dotted line shows ± two standard deviations.

Figure 4b shows the corresponding data and models for depth to the footwall. As noted earlier, the lack of a clear geological distinction at the footwall makes the wireframe a less reliable predictor of the footwall location than of the hangingwall location — $\Delta D(\mathbf{u})$ values are generally larger in magnitude on the footwall side.

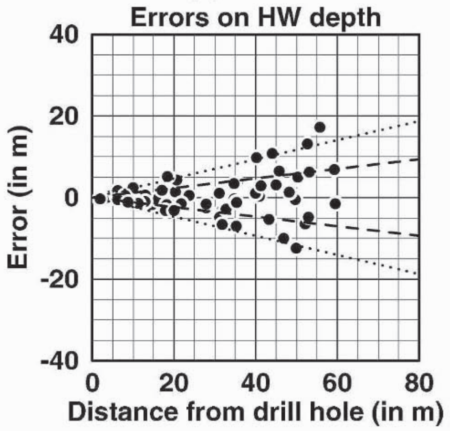


Figure 4a. $\Delta D(\mathbf{u})$ versus distance from nearest drill hole for hangingwall.

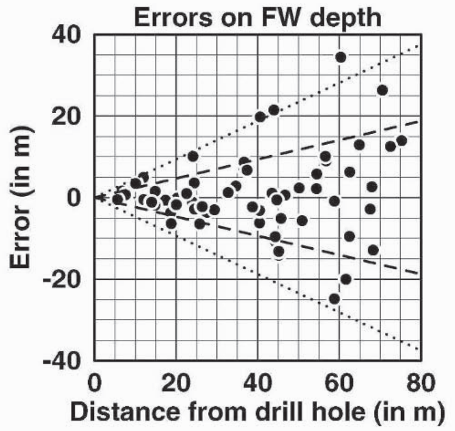


Figure 4b. $\Delta D(\mathbf{u})$ versus distance from nearest drill hole for footwall.

Using the dashed lines in Figure 4a and 4b, local cdfs of $\Delta D(\mathbf{u})$ can be constructed at every point on the mineralized surface. The distance from each point on the surface to the nearest drill hole intercept is calculated. Reading up from the x-axis on Figure 4 to the dashed line and across to the y-axis gives the standard deviation of $\Delta D(\mathbf{u})$ at that location; the mean is assumed to be zero and the shape of the cdf is assumed to be Gaussian. Figure 5 shows the median and $\pm 2\sigma$ bands of the local cdfs for the section shown in Figure 3.

With the local cdfs now established, all that remains is to sample them using a P-field with an appropriate range of

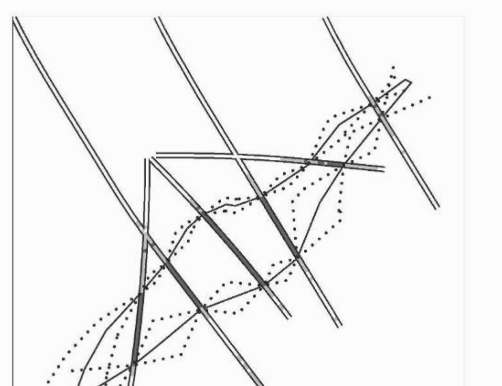


Figure 5. Local cdfs shown as a $\pm 2\sigma$ band around the median.

spatial autocorrelation. Figure 6 shows variograms of the uniform transform of the 62 $\Delta D(\mathbf{u})$ values from the second drilling campaign, along with their variogram models. Using these variograms, two 2D fields of spatially correlated probability values were created, one for use on the hangingwall and one for use on the footwall; these autocorrelated p values were then used to sample the local cdfs. Two of the resulting 100 realizations are shown in Figure 7.

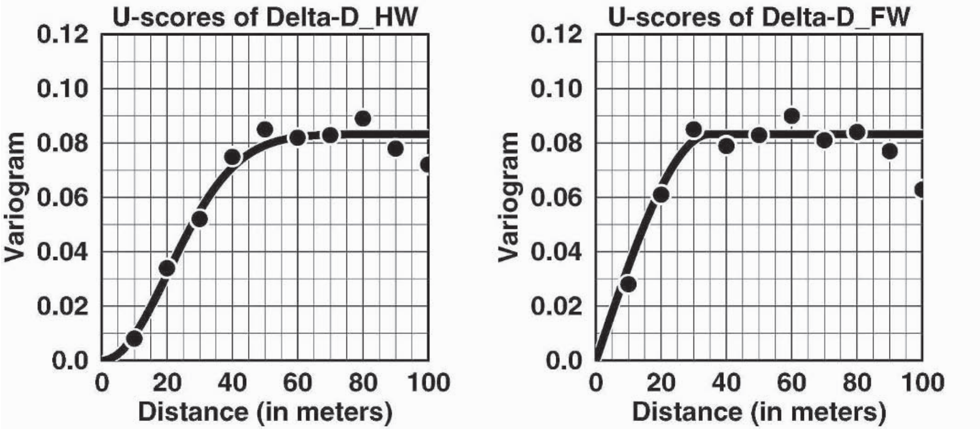


Figure 6a. Variogram for $\Delta D(\mathbf{u})$ on the hangingwall contact.

Figure 6b. Variogram for $\Delta D(\mathbf{u})$ on the footwall contact.

This example highlights the fact that local cdfs need not be estimated by kriging. In this case, they are established instead by a straightforward calibration based on historical data. This example also illustrates that the theoretical complexities of the P-field's statistical properties need not be an impediment to practical application. The uniform scores provide an experimental variogram that is easily modelled and that, when used to create unconditional P-fields, leads to geologically plausible results that greatly assist the assessment of project uncertainty and risk.

6 Conclusions

Even with exponential advances in computational speed, and the availability of many newer simulation algorithms, P-field simulation will likely remain one of the most often used geostatistical simulation procedures. Whenever local cdfs are already available and do not need to be generated using kriging, the P-field approach will be attractive since it decouples the issue of estimating cdfs from the task of sampling them. This not only reduces computational overhead, it also allows the user to generate realizations that fluctuate around a predetermined "base case", an advantage in many resource-based studies where a best estimate of global resources has already been established.

For studies that call for rapid generation of large numbers of conditional realizations, P-field will be attractive for its computational speed. Even as computer power has made it possible to run simulations hundreds of times faster than ten years ago, the appetite for larger simulations and for more realizations has kept

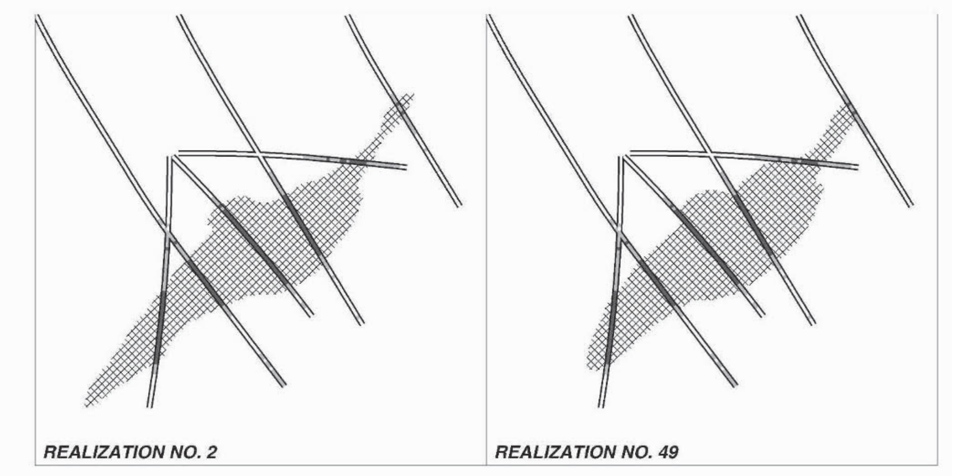


Figure 7. Two realization of the mineralized wireframe

pace. When a few realizations containing millions of grid nodes were once satisfactory, it is now not uncommon to generate hundreds of realizations containing tens of millions of grid nodes.

At the same time that P-field simulation will continue to be a good choice for many common applications, it is clear that there are many other common applications for which it is not a good choice. In particular, its tendency to create local extremes at conditioning locations makes it undesirable whenever downstream use of the realizations involves post-processing that is influenced by such artifacts. Fluid flow and contaminant transport studies are two examples of applications in which permeability extremes at wells or boreholes are clearly undesirable.

References

Cassiraga, E.F., 1999, *Incorporation de información blanda para la cuantificación de la incertitumbre: aplicación a la hidrogeologia*, Ph.D. Thesis, Universidad Politecnica de Valencia, Spain.

Froidevaux, R., 1993, "Probability Field Simulation", *Geostatistics Troia 1992*, A. Soares (ed.), Volume 1, Kluwer Academic Publishers, Dodrecht, Holland, pp. 73–84.

Goovaerts, P., "Geostatistical modelling of spatial uncertainty using p-field simulation with conditional probability fields", *International Journal of Geographical Information Sciences*, vol. 16, no. 2, pp. 167–178.

Journel, A.G., 1995, "Probability fields: another look and a proof", *SCRF Annual Report 8*, Stanford Center for Reservoir Forecasting, Stanford University, U.S.A.

Journel, A.G., and Ying, Z., 2001, "The theoretical links between sequential Gaussian simulation, Gaussian truncated simulation and probability field simulation, *Mathematical Geology*, vol. 33, no. 1, pp. 31–40.

Luster, G. R., 1985, *Raw Materials for Portland Cement: Applications of Conditional Simulation of Coregionalization*, Ph.D. Thesis, Stanford University, U.S.A.

Pyrcz, M., and Deutsch, C.V., 2001, "Two artifacts of probability field simulation", *Mathematical Geology*, vol. 33, no. 7, pp. 775–800.

Srivastava, R.M., 1992, "Reservoir characterization with probability field simulation", *SPE Paper No. 24753*, SPE Annual Conference and Exhibition, Washington, DC, U.S.A.

Tran, T., 1994, "Improving variogram reproduction on dense simulation grids", *Computers and Geosciences*, vol. 20, pp.1161–1168.

SEQUENTIAL SPATIAL SIMULATION USING LATIN HYPERCUBE SAMPLING

PHAEDON C. KYRIAKIDIS
Department of Geography, University of California Santa Barbara, Ellison Hall 5710, Santa Barbara, CA, 93106-4060, U.S.A.

Abstract. An efficient method is proposed for generating realizations from an arbitrary multivariate distribution using sequential simulation and Latin hypercube sampling. In a spatial context, this efficiency entails a reduction of sampling variability in statistics of spatially distributed model outputs when the inputs are realizations of random field models. The proposed method yields an unbiased reproduction of a target semivariogram, even for a small number of realizations, and consequently can be used for enhanced uncertainty and sensitivity analysis in complex spatially distributed models. In addition, the method is simple enough to be incorporated in virtually any geostatistical software for sequential simulation.

1 Introduction

Monte Carlo simulation is routinely used for uncertainty and sensitivity analysis of model outputs in a wide spectrum of scientific disciplines (Morgan and Henrion, 1990). Any realistic uncertainty analysis, however, calls for the availability of a representative distribution of such outputs, and can become extremely expensive in terms of both time and computer resources in the case of complex models and simple random (SR) sampling. This problem is far more pronounced for spatially distributed models, due to the large number of correlated (regionalized) variables comprising each input parameter map to such models, e.g., 3D rasters of hydraulic conductivity used for simulation of flow and transport in porous media.

An intelligent alternative to SR sampling is Latin hypercube (LH) sampling, a special case of stratified random sampling, which yields a more representative distribution of model outputs (in terms of smaller sampling variability of their statistics) for the same number of input simulated realizations. Analytical results demonstrating the efficiency of LH over SR sampling from univariate distributions are given in the (now classic) paper of McKay et al. (1979). A more recent comprehensive review of LH sampling for uncertainty and sensitivity analysis in complex systems can be found in Helton and Davis (2003).

LH sampling from a multivariate distribution, i.e., the task of inducing correlation in LH samples, is an important research theme in risk analysis and reliability

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 65-74.

engineering (Haas, 1999), which becomes critical in a spatial context for ensuring unbiased outputs of complex spatially distributed models. This paper makes a novel contribution to the literature of spatial uncertainty analysis, by proposing a simple and efficient method for sequential LH sampling from random field models.

2 Latin hypercube sampling

Consider a set of K independent continuous RVs $\{Y_k, k = 1, \dots, K\}$, with $F_{Y_k}(y_k) = Prob\{Y_k \leq y_k)$ denoting the cumulative distribution function (CDF) of the k-th RV Y_k. Simple random (SR) sampling of N realizations from RV Y_k proceeds by first generating a $(N \times 1)$ vector $\mathbf{u}_k = [u_k^{(n)}, n = 1, \dots, N]'$ of uniform random numbers in $[0, 1]$, which are treated as simulated probability values, and then transforming $\mathbf{u}_k$ into a $(N \times 1)$ vector $\mathbf{y}_k = [y_k^{(n)}, n = 1, \dots, N]$ of simulated realizations as: $\mathbf{y}_k = F_{Y_k}^{-1}(\mathbf{u}_k)$, using the inverse CDF $F_{Y_k}^{-1}$ of RV Y_k.

Latin hypercube (LH) sampling of N realizations from the k-th RV Y_k calls for generating, independent of vector $\mathbf{u}_k$, a $(N \times 1)$ vector $\mathbf{p}_k = [p_k^{(n)}, n = 1, \dots, N]'$ of random permutations of N integers $\{1, 2, \dots, N\}$. A $(N \times 1)$ vector $\mathbf{z}_k = [z_k^{(n)}, n = 1, \dots, N]'$ of stratified realizations is then obtained as (McKay et al., 1979):

$$\mathbf{z}_k = F_{Y_k}^{-1}\left(\frac{\mathbf{p}_k - \mathbf{u}_k}{N}\right) \tag{1}$$

where the argument $(\mathbf{p}_k - \mathbf{u}_k)/N$ of the inverse CDF $F_{Y_k}^{-1}$ ensures that the simulated probability values for the k-th RV Y_k are stratified, i.e., fall in N different probability strata. The monotonic transformation of the simulated probabilities incurred by the inverse CDF $F_{Y_k}^{-1}$ does not ruin stratification, which entails that each entry of vector $\mathbf{z}_k$ (each simulated value) falls within a different stratum in the original variable space, no matter the distributional form of $F_{Y_k}(y_k)$. The independence of vectors $\mathbf{p}_k$ and $\mathbf{u}_k$ ensures that there is a uniform probability $1/N$ for a simulated value within a particular stratum, i.e., there is no systematic placement of simulated values at the edges of strata. Variations of the above basic LH sampling procedure to further control sampling variability include variance reduction techniques, such as antithetic and control variates, as well as correlated sampling (Ang and Tang, 1984; Switzer, 2000).

A naive application of the above LH sampling procedure to correlated RVs fails to induce any correlation in the simulated values, simply because vectors $\mathbf{p}_k$ and $\mathbf{u}_k$ for the k-th RV Y_k are generated independent of other such vectors for other RVs. From these two sources that contribute to lack of correlation, the most important one is the vector $\mathbf{p}_k$ of random permutations because it dictates the strata within which the entries of $\mathbf{u}_k$ are distributed. To date, the most widely used method for generating LH samples from correlated RVs with a given rank correlation coefficient is the distribution-free method of Iman and Conover (1982). This method, however, is prohibitive for a large number $K > 10,000$ of RVs (typically the case in a spatial setting) because it calls for the Cholesky decomposition of an extremely large $(K \times K)$ variance-covariance matrix.

Stein (1987) proposed a (now also widely used) post-processing method for transforming a SR sample from K correlated RVs into a LH sample. Stein's method is independent of the simulation algorithm used to generate the original SR sample, and can be applied in principle to a large number K of RVs. Let $\mathbf{Y} = [\mathbf{y}_k, k = 1, \ldots, K]$ denote a $(N \times K)$ matrix containing a SR sample of size N from the K-variate CDF of the above K RVs; the k-th column $\mathbf{y}_k$ of this matrix corresponds to outcomes of the k-th RV Y_k. Matrix $\mathbf{Y}$ can be generated, for example, by simulation via the Cholesky decomposition of the covariance matrix, or via sequential simulation (Johnson, 1987). The SR sample $\mathbf{y}_k$ for the k-th RV Y_k is then transformed into a LH sample $\mathbf{z}_k$ for that RV, as:

$$\mathbf{z}_k = F_{Y_k}^{-1}\left(\frac{\mathbf{r}_k - \mathbf{u}_k}{N}\right) \tag{2}$$

where $\mathbf{r}_k = [r_k^{(n)}, n = 1, \ldots, N]'$ denotes a $(N \times 1)$ vector containing the ranks of the entries of $\mathbf{y}_k$: the lowest $y_k^{(n)}$ simulated value for the k-th RV Y_k is assigned a rank of one, the second lowest a rank of two, and the highest a rank of N.

Stein's method is similar to the LH sampling method of Equation (1), with the sole, but extremely important, difference that the array $\mathbf{p}_k$ of random permutations in that equation is now replaced by the array $\mathbf{r}_k$ of ranks of $\mathbf{y}_k$. This substitution entails that the LH sample comprising the $(N \times K)$ matrix $\mathbf{Z} = [\mathbf{z}_k, k = 1, \ldots, K]$ is (column-wise) correlated, since it inherits correlation that is present in the SR sample $\mathbf{Y}$ via the corresponding $(N \times K)$ matrix $\mathbf{R} = [\mathbf{r}_k, k = 1, \ldots, K]$ of its ranks. In addition, the entries of any column of matrix $\mathbf{Z}$ are stratified, as opposed to the entries of any column of matrix $\mathbf{Y}$.

Figure 1 gives an example of a SR sample (A) and a LH sample generated using Stein's method (B), both of size $N = 10$, from two standard Gaussian RVs Y_1 and Y_2 with correlation coefficient $\rho_{12} = 0.7$. It can be easily appreciated that, for the LH sampling case, realizations for both RVs are marginally stratified, i.e., when viewed from either the abscissa or the ordinate, each stratum (delineated by vertical or horizontal solid lines, respectively) contains a single simulated value.

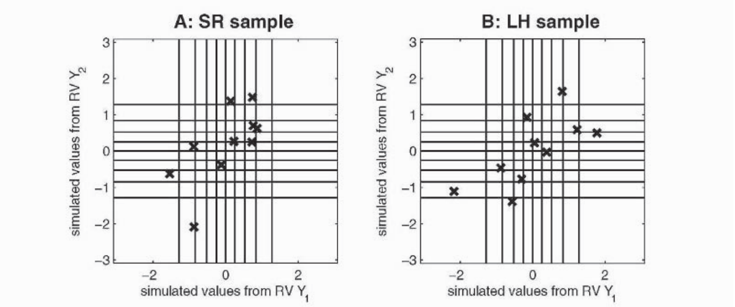


Figure 1. Examples of a SR sample (A), and a LH sample generated using Stein's method (B), both of size $N = 10$, from two correlated standard Gaussian RVs Y_1 and Y_2 with $\rho_{12} = 0.7$; solid lines delineate strata of equal probability.

Stein's method, however, underestimates the target correlation between any two RVs, because it does not fully account for the correlation in the original SR sample. More precisely, the sole vehicle for inducing correlation in the LH sample $\mathbf{z}_k$ for RV Y_k is the rank vector $\mathbf{r}_k$ of the original SR sample $\mathbf{y}_k$ for that RV; see Equation (2). The vector $\mathbf{u}_k$ of uniform random numbers in that equation is generated independent from any other such vector $\mathbf{u}_{k'}$ for any other RV $Y_{k'}$. For small sample sizes (small N) the displacement in the probability axis of the original uniform random vector that generated $\mathbf{y}_k$ (brought by the new vector $\mathbf{u}_k$) can be large; this affects the reproduction of a target correlation by the LH sample.

The above underestimation of a target correlation was also corroborated empirically in a spatial setting by Pebesma and Heuvelink (1999), who applied Stein's post-processing method to transform a SR sample generated via sequential Gaussian simulation to a LH sample. Their results showed that simulated realizations exhibited small-scale variability larger than that dictated by the target semivariogram model. This bias was also shown to be higher for small sample sizes, which unfortunately is precisely the reason for employing LH sampling in the first place.

In what follows, Stein's method is adopted not as a post-processing step, but as an integral part of sequential simulation for generating a LH sample from a multivariate distribution. To the author's knowledge, the proposed LH sampling method constitutes a novel contribution to the literature of importance sampling.

3 Sequential Latin hypercube sampling

Let $F_{Y_1,\ldots,Y_K}(y_1,\ldots,y_K|\mathbf{d}) = Prob\{Y_1 \leq y_1,\ldots,Y_K \leq y_K|\mathbf{d}\}$ denote the K-variate conditional CDF (CCDF) of K RVs $\{Y_k, k=1,\ldots,K\}$, given a $(O \times 1)$ vector $\mathbf{d} = [d_o, o=1,\ldots,O]'$ with known realizations (sample observations) of O RVs $\{Y_o, o=1,\ldots,O\}$. Conditional stochastic simulation amounts to generating N alternative realizations from the multivariate CCDF $F_{Y_1,\ldots,Y_K}(y_1,\ldots,y_K|\mathbf{d})$, whereas unconditional simulation corresponds to absence of sample observations, in which case the data vector $\mathbf{d}$ is simply dropped from the notation.

The multiplication rule of probability allows one to decompose the above K-variate CCDF into a sequence of K univariate CCDFs as:

$$F_{Y_K,\ldots,Y_1}(y_K,\ldots,y_1|\mathbf{d}) = F_{Y_K}(y_K|y_{K-1},\ldots,y_2,y_1,\mathbf{d})\cdots F_{Y_2}(y_2|y_1,\mathbf{d})F_{Y_1}(y_1|\mathbf{d}) \tag{3}$$

which entails that the n-th SR sample from the above multivariate CCDF can be generated sequentially by first simulating a value $y_1^{(n)}$ from CCDF $F_{Y_1}(y_1|\mathbf{d})$, then a simulated value $y_2^{(n)}$ from CCDF $F_{Y_2}(y_2|y_1^{(n)},\mathbf{d})$, and so forth.

It is important to note that all the above univariate CCDFs, apart from the first one $F_{Y_1}(y_1|\mathbf{d})$, change from one realization to another, because the previously simulated values used as conditioning data are different for each realization. The CCDF of the k-th RV Y_k for the n-th realization should thus be denoted as: $F_{Y_k}^{(n)}(y_k|\mathbf{y}_{k-1}^{(n)},\mathbf{d})$, where $\mathbf{y}_{k-1}^{(n)} = [y_l^{(n)}, l=1,\ldots,k-1]$ is the $(1 \times k-1)$ vector of simulated values generated prior to $y_k^{(n)}$. In expected value (over a large number N of realizations), however, the CCDF for any RV Y_k tends towards its CCDF given only the data vector $\mathbf{d}$, i.e., $E\{F_{Y_k}(y_k|\mathbf{Y}_{k-1}, \mathbf{d})\} \simeq F_{Y_k}(y_k|\mathbf{d})$, where $\mathbf{Y}_{k-1} = [\mathbf{y}_l, l = 1, \ldots, k-1]$ is the $(N \times k-1)$ matrix of all simulated values for all RVs considered before Y_k in all N realizations.

The proposed LH sampling method from an arbitrary multivariate distribution capitalizes on the above decomposition, and amounts to embedding Stein's method into sequential simulation, which now proceeds in the following steps:

1. Establish a sequence for considering all K RVs. As long as all simulated values generated from any RV in this sequence are used as conditioning information (in addition to the data vector $\mathbf{d}$) for simulation from subsequent RVs, the order of the sequence is irrelevant: the resulting realizations constitute a genuine sample from the multivariate CCDF of Equation (3).
2. For the k-th RV Y_k in the above sequence:
   a) establish all N CCDFs $\{F^{(n)}_{Y_k}(y_k|\mathbf{Z}_{k-1}, \mathbf{d}), n = 1, \ldots, N\}$, each corresponding to a particular realization n; $\mathbf{Z}_{k-1}$ is a $(N \times k-1)$ matrix with the entire LH sample generated in all N realizations before considering RV Y_k.
   b) generate a $(N \times 1)$ vector $\mathbf{y}_k$ with a SR sample from RV Y_k; the n-th entry $y^{(n)}_k$ of vector $\mathbf{y}_k$ is drawn from the n-th CCDF $F^{(n)}_{Y_k}(y_k|\mathbf{z}^{(n)}_{k-1}, \mathbf{d})$, where $\mathbf{z}^{(n)}_{k-1}$ is a $(1 \times k-1)$ vector with the n-th LH sample generated from all RVs considered before Y_k, i.e., $\mathbf{z}^{(n)}_{k-1}$ is the n-th row of matrix $\mathbf{Z}_{k-1}$.
   c) transform the SR sample $\mathbf{y}_k$ into a LH sample $\mathbf{z}_k$, as:

   $$\mathbf{z}_k = F^{-1}_{Y_k|\mathbf{d}}\left(\frac{\mathbf{r}_k - \mathbf{u}_k}{N}\right) \tag{4}$$

   where $F^{-1}_{Y_k|\mathbf{d}}$ denotes the inverse CCDF of RV Y_k given only the data vector $\mathbf{d}$, $\mathbf{r}_k$ is the rank transform of $\mathbf{y}_k$, and $\mathbf{u}_k$ is a vector of uniform random numbers in $[0, 1]$ (independent of $\mathbf{y}_k$).
   d) augment the LH sample matrix $\mathbf{Z}_{k-1}$ of step 2a to obtain the current LH sample matrix $\mathbf{Z}_k = [\mathbf{Z}_{k-1}\ \mathbf{z}_k]$ of size $(N \times k)$.
3. Consider the next RV Y_{k+1} in the sequence established in step 1, and repeat step 2 for generating LH samples from all remaining RVs $\{Y_l, l = k+1, \ldots, K\}$.

In the proposed approach, the LH sampling method of Stein is used as a post-processing tool (step 2c) *after* drawing a SR sample $\mathbf{y}_k$ from the N univariate CCDFs of RV Y_k (step 2b). But, unlike Stein's method, the LH sample $\mathbf{z}_k$ for RV Y_k is generated *before* proceeding to the simulation of the next SR sample $\mathbf{y}_{k+1}$ from the subsequent RV Y_{k+1} (step 3). Most importantly, that LH sample $\mathbf{z}_k$ is also considered as conditioning information for simulation from all subsequent RVs $\{Y_l, l = k+1, \ldots, K\}$ (step 2d), which leads to the reproduction of a target (conditional) correlation per the theory of sequential simulation (Journel, 1994).

In principle, any linear or non-linear regression scheme can be used to determine the CCDF of any RV Y_k (step 2a); the proposed LH sampling method, however, is independent of the particular scheme adopted for this CCDF determination. When the multivariate CCDF of Equation (3) is Gaussian, the CCDF of any RV Y_k (step 2a) is univariate Gaussian, and thus fully characterized by its conditional mean and variance which can be derived via generalized linear regression (Kriging); this is also the building block of sequential Gaussian simulation in a spatial context (Deutsch and Journel, 1998). LH samples from non-Gaussian RVs with specified pairwise rank correlations can also be generated by first simulating correlated deviates from K Gaussian RVs, and then transforming these deviates to correlated realizations of the original RVs using the inverse marginal or conditional CDF of each RV (Iman and Conover, 1982).

Since an unbiased (in expected value) reproduction of a target correlation is only ensured in sequential simulation under SR sampling, a hybrid approach between LH and SR sampling (still in a sequential mode) is also investigated in this paper. More precisely, this second proposal amounts to transforming the LH sample $\mathbf{z}_k$ for RV Y_k (step 2c above) to a new LH sample $\mathbf{x}_k$ that is as close as possible to the corresponding SR sample $\mathbf{y}_k$ for that RV, under the constrain that the elements of this new sample $\mathbf{x}_k$ remain in the strata used in the LH sampling procedure. In other words, the elements of the original LH sample $\mathbf{z}_k$ are "displaced" *within their strata* towards the corresponding elements of the SR sample $\mathbf{y}_k$ with the same rank. In the remainder of this paper, SRS denotes simple random sampling, LHSS denotes the LH sampling method of Stein, LHSP1 denotes the first proposal for LH sampling outlined in the flowchart given above, and LHSP2 denotes this second proposal for hybrid LH sampling.

Figure 2 gives the sampling distributions of correlation coefficients calculated from 10000 sets of LH samples, each of size $N = 10$, generated from two standard Gaussian RVs Y_1 and Y_2 with correlation $\rho_{12} = 0.8$ using the four sampling methods considered in this work. In this case, no data vector $\mathbf{d}$ is considered (unconditional simulation), and the CCDFs of RV Y_2 given realizations of RV Y_1 (step 2a) are determined via simple Kriging. The unbiased reproduction of the target correlation from SRS and LHSP2 (Figures 2A and D) is evident. Both LHSS and LHSP1 exhibit a bias in the reproduction of the target correlation. For LHSS (Figure 2B) that bias is -6%, whereas for LHSP1 (Figure 2C) it is reduced to -2%. Note that any bias decreases for larger sample sizes (larger N values).

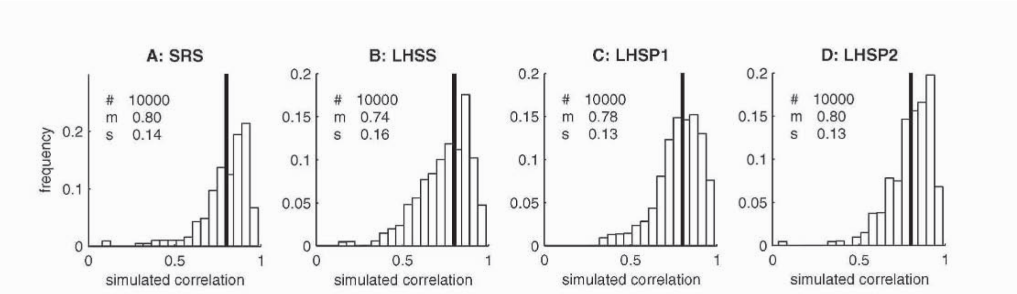


Figure 2. Sampling distributions of correlation coefficients calculated from 10000 sets of simulated pairs (each set of size $N = 10$) generated from two correlated standard Gaussian RVs Y_1 and Y_2 via: SRS (A), LHSS (B), LHSP1 (C), and LHSP2 (D); solid lines indicate the target correlation coefficient $\rho_{12} = 0.8$.

4 Spatial Latin hypercube sampling

In a spatial setting, when all K RVs pertain to a single spatial attribute Y (univariate case), the k-th RV $Y_k = Y(\mathbf{s}_k)$ is defined at a location with coordinate vector $\mathbf{s}_k$. The o-th entry $d(\mathbf{s}_o)$ of the data vector $\mathbf{d}$ denotes the sample attribute value at the o-th observation site with coordinate vector $\mathbf{s}_o$. The objective is then to generate simulated realizations (typically up to 3D) from the multivariate distribution of Equation (3), conditional or not on the data vector $\mathbf{d}$. Different sequential spatial simulation methods can be distinguished according to how each univariate CCDF is determined at each location (step 2a). Variants of Kriging are typically used for building such local CCDFs (Deutsch and Journel, 1998; Chilès and Delfiner, 1999), or more recently multi-point statistics when training images are available (Strebelle, 2000).

Sequential spatial simulation typically proceeds on a random path (different from one realization to another) for visiting each simulation location. This avoids the creation of artifact patterns in the realizations, when not all previously simulated values are used as conditioning information at any location along this path (Deutsch and Journel, 1998). In the proposed approach, that path can also be random, but it must be the *same* for all realizations (step 1); in any other case, a LH sample can only be obtained *after* sequential simulation, using the original method of Stein (1987) with its shortcomings for small N. In the examples of SR and LH sampling of this paper, a *single* random path is considered, and *all* previously simulated values are used as conditioning data at any simulation grid node to eliminate the impact of different search strategies on sampling variability.

The reproduction of target statistics from the four sampling methods considered was initially investigated using a single sample of $N = 10$ realizations generated via unconditional sequential Gaussian simulation at $K = 300$ nodes of a regular unit-spaced 1D grid. The stationary statistics included a zero mean, a unit variance, and a spherical semivariogram model of range 30 distance units. Figure 3 gives the reproduction of the marginal mean and variance at each grid node. Evidently, all LH sampling methods lead to a significant reduction in sampling variability with respect to SR sampling (compare Figure 3A with Figures 3B-D).

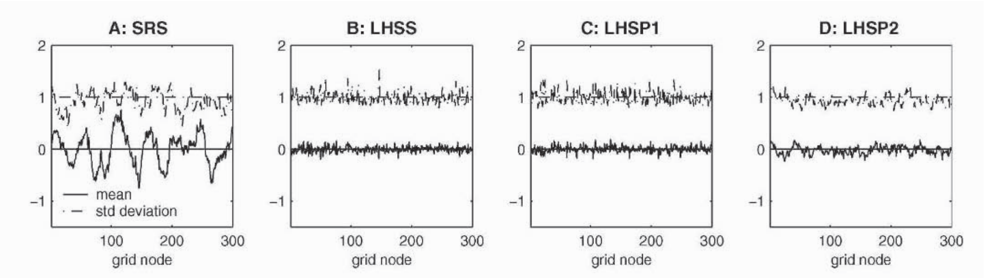


Figure 3. Mean and standard deviation of $N = 10$ simulated values at $K = 300$ unit-spaced nodes of a regular 1D grid, generated using unconditional sequential Gaussian simulation with: SRS (A), LHSS (B), LHSP1 (C), and LHSP2 (D). Horizontal solid and dashed-dotted lines indicate the target mean (0) and standard deviation (1), respectively.

Figure 4 gives the semivariogram reproduction of the $N = 10$ realizations, i.e., of the single sample whose marginal statistics are shown in Figure 3. Stein's LH sampling method (Figure 4B) leads to a higher mean simulated semivariogram than the target model at small lag distances; this critical underestimation of spatial correlation is significantly reduced by the original proposal LHSP1 (Figure 4C), and virtually eliminated by the hybrid proposal LHSP2 (Figure 4D). As expected, SR sampling leads to an unbiased mean simulated semivariogram, especially at the critical small lag distances (Figure 4A). Note also the somewhat smaller variance of the simulated semivariogram for LHSS and LHSP1 than for SR sampling (for this particular sample).

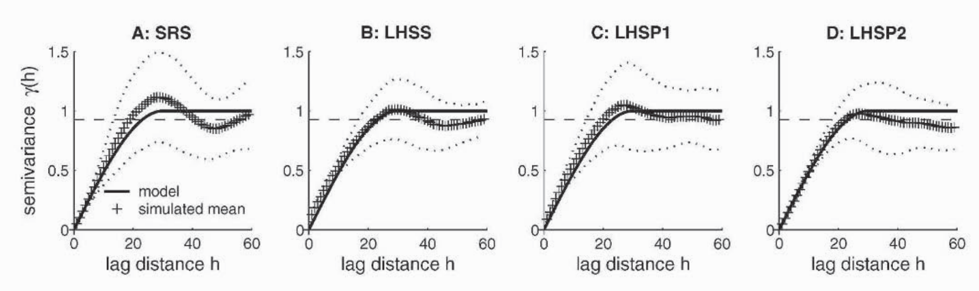


Figure 4. Semivariogram reproduction from a single sample of $N = 10$ simulated realizations at $K = 300$ unit-spaced nodes of a regular 1D grid, generated using unconditional sequential Gaussian simulation with: SRS (A), LHSS (B), LHSP1 (C), and LHSP2 (D). Solid lines indicate the target, unit-sill, spherical semivariogram model with range 30 distance units; crosses indicate the mean simulated semivariance at each lag; dotted lines delineate intervals of one standard deviation from either side of the mean semivariance; dashed horizontal lines indicate the average semivariogram value $\bar{\gamma}(V, V) = 0.972$, where V denotes the line segment support of 300 units.

To further assess the efficiency of spatial LH sampling, 1000 independent sets of $N = 10$ realizations were generated at $K = 100$ unit-spaced nodes of a regular 1D grid, using unconditional sequential Gaussian simulation and the four sampling methods considered in this work. The semivariogram model adopted was a unit-sill spherical semivariogram of 30 distance units; a stationary zero mean was also assumed. The statistic under consideration is the proportion of simulated values above threshold $G^{-1}(0.75) = 0.6745$, when these values are arranged in groups of three or more contiguous ("connected") nodes; here G^{-1} denotes the inverse Gaussian CDF. This latter connectivity consideration allows evaluating any bias incurred by a poor semivariogram reproduction: a larger than expected nugget effect, for example, will lead to a smaller than expected number of connected groups containing at least three nodes. The reference proportion 0.2219 of such connected nodes was established from a large SR sample of size $N = 1000$.

Figure 5 gives the sampling distribution of the simulated mean proportions for the four sampling methods considered in this experiment. The significant reduction in sampling variability incurred by the LH sampling methods with respect to SR sampling is evident (compare Figure 5A to Figures 5B-D). Such a reduction, however, comes at the expense of a bias in the case of Stein's method (Figure 5B); that bias is almost absent from the results of the proposed methods (Figures 5C-D).

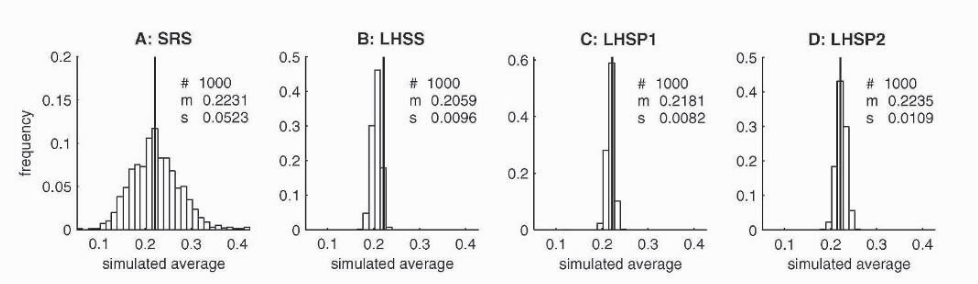


Figure 5. Sampling distributions of the mean proportion of "connected" simulated values above threshold 0.6745, for SRS (A), LHSS (B), LHSP1 (C), and LHSP2 (D); see text for details. Solid lines indicate the target proportion of 0.2219, calculated from a large SR sample of size $N = 1000$.

5 Discussion and conclusions

A novel method for LH sampling from random field models in a sequential mode has been presented in this paper. The original proposal consists of transforming a SR sample to a LH sample at each step of sequential simulation using Stein's method. A further improvement consists of additional "displacements" of the elements of the LH sample for a particular variable, *within their respective strata*, towards the corresponding elements of the SR sample with the same rank. It has been demonstrated that both proposals significantly reduce sampling variability in resulting marginal statistics, and thus make better use of the same number of realizations than SR sampling. The main advantage of these proposals over the comparable method of Stein is their better (less biased) reproduction of a target semivariogram at small lag distances, even from few realizations, a critical requirement in a spatial context to ensure unbiasedness of model outputs.

It should be noted that LH sampling leads to a smaller sampling variability in statistics of model outputs, when these models are monotonic in their inputs; such a reduction is also larger for linear models (McKay et al., 1979; Stein, 1987). It is therefore important that target marginal statistics be correctly estimated. If deemed necessary, uncertainty in these statistics should be incorporated in a formal Bayesian framework, rather than via ergodic fluctuations of SR sampling.

The proposed LH sampling method is not limited to Gaussian random field models, continuous variables, point support values, or two-point statistics, because it is independent of the algorithm used to determine the local CCDFs in sequential simulation. Any practical approximation in the implementation of sequential simulation, such as the consideration of a limited number of previously simulated values at each location, is nevertheless shared by the proposed LH sampling method. The impact of this latter approximation is typically alleviated via cascaded simulation on nested grids of increasing resolution (Deutsch and Journel, 1998). Moreover, in many cases, e.g., for simulation from auto-regressive processes, sequential simulation is the natural way to generate realizations from such processes. When the number of simulation locations is very large, and such locations do not lie on a regular grid, sequential simulation is perhaps the only feasible algorithm, due to precisely its practical implementation approximations.

Concluding, the proposed sequential method for spatial LH sampling can be readily used for enhanced uncertainty and sensitivity analysis, as well as subsequent risk assessment, in situations where complex spatially distributed models are involved. In addition, the method is simple enough to be incorporated in virtually any geostatistical software for sequential simulation, and can handle a very large number of simulation locations.

Acknowledgements

Funding for this research was partially provided by the National Science Foundation under Award BCS #0322349.

References

Ang, A.H-S., and Tang, W.H. (1984): *Probability Concepts in Engineering Planning and Design – Volume II: Decision, Risk, and Reliability*, John Wiley & Sons.

Chilès, J.P., and Delfiner, P. (1999): *Geostatistics: Modeling Spatial Uncertainty*, John Wiley & Sons.

Deutsch, C.V., and Journel, A.G. (1998): *GSLIB: Geostatistical Software Library and User's Guide, 2nd Edition*, Oxford University Press.

Haas, C.N. (1999): On modeling correlated random variables in risk assessment, *Risk Analysis*, 19(6), 1205-1214.

Helton, J.C., and Davis, F.J. (2003): Latin hypercube sampling and the propagation of uncertainty in analyses of complex systems, *Reliability Engineering & System Safety*, 81(1), 23-69.

Iman, R.L., and Conover, W.J. (1982): A distribution-free approach to inducing rank correlation among input variables, *Communications in Statistics, Part B–Simulation and Computation*, 11(3), 311-334.

Johnson, M.E. (1987): *Multivariate Statistical Simulation*, John Wiley & Sons.

Journel, A.G. (1994): Modeling spatial uncertainty: Some conceptual thoughts, *in: Geostatistics for the Next Century*, R. Dimitrakopoulos (Ed.), p. 30-43, Kluwer Academic Publishers.

McKay, M.D., Beckman, R.J., and Conover, W.J. (1979): A comparison of three methods for selecting values of input variables in the analysis of output from a computer code, *Technometrics*, 21(2), 239-245.

Morgan, M.G., and Henrion, M. (1990): *Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis*, Cambridge University Press.

Pebesma, E.J., and Heuvelink, G.B.M. (1999): Latin hypercube sampling of Gaussian random fields, *Technometrics*, 41(4), 303-312.

Stein, M. (1987): Large sample properties of simulations using Latin hypercube sampling, *Technometrics*, 29(2), 143-151.

Strebelle, S. (2000): Conditioning simulation of complex geological structures using multi-point statistics, *Mathematical Geology*, 34(1), 1-21.

Switzer, P. (2000): Multiple simulation of spatial fields, *in: Proceedings of the 4th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences*, G.B.M. Heuvelink, and M.J.P.M. Lemmens (Eds.), p. 629-635, Coronet Books Inc.

FIELD SCALE STOCHASTIC MODELING OF FRACTURE NETWORKS -
Combining pattern statistics with geomechanical criteria for fracture growth

XIAOHUAN LIU and SANJAY SRINIVASAN
Department of Petroleum & Geosystems Engineering
University of Texas at Austin, United States, 78712-0228

Abstract: According to recent estimates, the U.S. domestic potential for fractured oil reservoirs is on the order of tens of billion of barrels. Better technology for characterizing fracture flow paths, especially in deep, non-conventional plays and in carbonate rocks is a key to producing hydrocarbons economically from these reservoirs. The paper presents an approach for stochastic, field-scale modelling of fracture networks consistent with patterns observed on logs, the physical basis for fracture propagation and field-specific observations.

1. Introduction

Two aspects of research are presented. A stochastic simulation approach that utilizes fracture pattern information retrieved from analog models is presented first. Pattern characteristics are inferred from outcrop images using multipoint statistics and subsequently applied, after affine transformations to simulate fracture patterns in the target reservoir. A unique, stochastic fracture growth-based simulation algorithm is presented for imposing the multipoint fracture pattern characteristics on the simulation models.

Fracture patterns observed in outcrops or in subsurface reservoirs can be explained in terms of the structural geology of the reservoir and spatial variations in mechanical properties of rocks. Fracture growth model based on geomechanics can be used to perform physics-based numerical simulation of fracture patterns. However, geomechanical models can only generate fracture patterns up to a length scale of 1 kilometer and the uncertainty in fracture characteristics due to uncertainty in the stress field cannot be quantified. The paper presents a multipoint-based approach to characterize fracture patterns inferred from geomechanical models and these statistics are merged with the pattern statistics inferred from analogs such as outcrop or logs information in order to generate field-scale reservoir models. A Bayesian approach for incorporating uncertainty in reservoir stress field is also presented. The probability of the stress field conditional on the observed pattern in well logs is calibrated using the geomechanical model and later inverted to yield the probability of a fracture pattern given the uncertainty in the reservoir stress levels. Finally fractures are propagated in the reservoir by applying the multiple-point simulation approach constrained to the previously derived probabilities.

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 75-84.

2. Geomechanical fracture classification and stochastic simulation

A natural fracture is a planar discontinuity in reservoir rock due to deformation or physical diagenesis[1]. Natural fracture patterns are frequently interpreted on the basis of laboratory-derived fracture patterns corresponding to models of paleo-stress fields and strain distribution in the reservoir at the time of fracture[2]. Sterns and Friedman[3] proposed a genetic classification of fracture systems based on stress/strain conditions in laboratory samples and features observed in outcrops and sub-surface settings. Based on their work, it can be concluded that complex stress and strain distributions in reservoir rocks can result in complex fracture patterns. Fracture patterns corresponding to different geological systems have key characteristics that can be used to classify and index fracture networks observed in outcrops and subsurface samples. Multiple point statistical measures can be used for identifying and classifying fracture patterns corresponding to different fracture systems[4].

Since stress boundary conditions strongly control the fracture pattern development subsurface at the time of fracturing, a geomechanics-based approach, where a physical understanding of the fracturing process is combined with measurements of mechanical properties of rock, is physically realistic to predict fracture network characteristics. This process-oriented approach can also provide a theoretical basis for deciding what types of fracture attribute distributions are physically reasonable, and how attributes such as length, spacing and aperture are inter-related. Additional geological information, such as the strain, pore pressure and diagenetic[5] history of the reservoir can provide further constraint on fracture network predictions.

In most cases, data available to model the fractured reservoir are sparse and information such as seismic maps and production response are related imprecisely to the fracture pattern characteristics, a probabilistic approach to fracture characterization is necessary. In the object-based modeling approaches, fractures are represented as objects defined by their centroid, shape, size and orientation. In “Random Disk” models[6], fractures are represented as two-dimensional convex circular disks located randomly in space. Although object-based models are easy to implement, their application is limited due to the assumed independence of the model parameters such as radii, orientation etc. A viable alternative is to employ pixel-based algorithms. Well established geostatistical algorithms such as sequential indicator simulation (sisim)[7] ensure reproduction of the two-point indicator variogram and can be used to classify nodes within the reservoir into fractures or matrix. However, models constrained only to two-point statistics are generally noisy and consequently inadequate for capturing clean-cut shapes such as fractures.

Although stochastic fracture models can be constructed that might be representative of analog fracture reservoir to some degree, it is still difficult to integrate geomechanical information such as stress boundary conditions in those models. A promising conditional multiple point simulation approach with integration of geomechanical data is developed as part of this research.

3. Multiple point approach to fracture growth simulation

3.1 Simulation algorithm

In the case of traditional two-point statistics based algorithms the cumulative conditional distribution function (CCDF) depicting the local uncertainty in attribute value is calculated on the basis of two-point correlation between pairs of data and between each data and simulation node. In multiple point statistics based algorithms[8, 9], this required conditional probability distribution is derived based on the entire data configuration on a spatial template, including the multiple-point interactions among the data and between the data and the unknown. Supposing there are n neighboring data events $A_\alpha, \alpha = 1,.....,n$. An additional variable $t(n) = 1$ is assigned if all the elementary data events occur simultaneously. The conditional probability is[9]:

$$\text{Prob}\{A_0 = 1|t(n) = 1\} = E\{A_0 = 1|t(n) = 1\} \tag{1}$$

A_0 is the unknown data at the unsampled location. Using Bayes' Theorem, the conditional probability in expression (1) can be written as:

$$\text{Prob}\ \{A_0 = 1\,|\,t(n) = 1\} = \frac{\text{Prob}\ \{A_0 = 1, t(n) = 1\}}{\text{Prob}\ \{t(n) = 1\}} \tag{2}$$

This implies that in order to derive the multiple-point conditional probability expression (1), we need to know the joint probability of observing the spatial pattern $A_0 = 1$ and $t(n) = 1$ as well as the prior probability of the occurrence of the template pattern $t(n) = 1$. Given an analog fracture model e.g. based on outcrop exposures, the required probabilities can be retrieved from that model. Defining a spatial template and translating that template over the analog model, the joint frequency of events such that $A_0 = 1$ and $t(n) = 1$ as well as the prior probability of events $t(n) = 1$ can be retrieved.

The fracture simulation approach adopted in this research exhibits a distinct departure from the current state-of-the-art multiple-point statistics based approaches in that the simulation event A_0 is itself considered to be a multiple-point event, obtained constrained to $t(n) = \cup A_\alpha$, a multiple-point event of arbitrary complexity. In contrast, in the traditional multiple-point simulation approaches, the simulation event A_0 is generally treated as single point event. As a consequence of this subtle and yet significant departure from other traditional methods, fractures are grown from each seed location based on the probability of the multiple point simulation events A_0 inferred from analog fracture models. The seed fracture locations are selected based on areal proportion maps that may be derived from seismic maps or other physical criteria such as surface curvature maps. Such a growth algorithm has the advantage that it is computationally efficient and permits integration of other physical criteria for fracture growth that might be controlled by variations in mechanical properties of the rock.

The simulation commences from an empty grid. The well locations with recorded fracture data are visited sequentially. The data configuration on a 27-point template

(Figure 1) surrounding the fracture location is examined. The conditioning data includes original well data as well as well nodes that have already simulated to be fractures. The analog fracture model is then scanned for the occurrence of that data configuration. Thus, if for example as in Figure 1, at the current stage of simulation, there are 23 points surrounding the central node that have been previously simulated to be fractures, then the analog model is scanned for the occurrence of that 24-point (23+1 central node) data configuration. This yields the probability $\text{Prob}\{t(n)=1\}$ corresponding to that data configuration. The simulation event A_0 can then be one of the following:

- None of the remaining three points on the template is a fracture
- One of the remaining three locations is a fracture. That location could be any one of the remaining nodes
- Two of the remaining three locations are fractures. There are three possible combinations.
- All three of the remaining three locations are also fractures.

The probability associated with all such multiple-point data events A_0 are retrieved by scanning the analog model. This is the joint probability $\text{Prob}\{A_0=1,t(n)=1\}$ corresponding to each data event A_0. The conditional probability: $\text{Prob}\{A_0=1,|\,t(n)=1\}$ is then derived as the ratio of the joint probability and the prior probability. A random value is drawn from the conditional probability distribution and this yields the set of nodes corresponding to the outcome A_0^* that are marked as fractures for the next step of the simulation algorithm.

3.2 Results discussion

As an example implementation of the simulation algorithm, Figure 2 is the training image of fracture distribution obtained as an unconditional realization of an object-based model. Since in most cases fracture patterns observed on a outcrop are on a 2-D plane, the analog model in Figure 2 as well as the subsequent multiple point simulation algorithm was implemented in 2-D. The spatial template used for retrieving the conditional probability distributions is shown in Figure 3. Figure 4 is the result of the simulation approach described above. It is easily to observe that the simulated model is consistent with the training image. For instance, there are three different fracture orientations (N-S, NE-SW, NW-SE) in training image (Figure 2) that can also be observed in simulation image; and both images exclude horizontal orientation fracture. It can be thus concluded that this multiple point statistical simulation approach can reproduce fracture patterns observed in training model/analog models.

4. Geomechanical Basis for Fracture growth

4.1 Principle of fracture growth

It is known that the observed strength of rocks in laboratory experiments is significantly lower than calculated theoretically values. This discrepancy is due to the remarkable strength reduction in rocks caused by stress concentrations at crack tips and subsequent propagation of pre-existing small flaws within rocks as well as other solid materials. The geomechanical modeling approaches discussed in research are based on the presence of pre-existing cracks known as Griffith cracks[5] within rocks. Stress concentrations and propagation will occur along cracks at an orientation consistent with the applied load. All modes (tensile, in-plane shear, anti-plane shear) of crack propagation with respect to different sizes, shapes and orientation of rocks and under various boundary conditions can be predicted by geomechanical analysis. Just as rocks have a critical tensile stress capacity[5], they also have a critical stress intensity factor K_c. The crucial criterion to propagate a crack through the rock is that the stress intensity at the crack tips be at least equal to the critical stress intensity. In long-term loading systems such as in petroleum reservoirs, classic fracture mechanics may fail to accurately predict the crack growth especially in the presence of high temperature and chemical reactivity. Crack propagation can thus occur at a stress intensity value K less than the critical intensity. This has been observed in experiments using many materials including rocks and minerals and is referred to as subcritical crack growth.

4.2 Description of Geomechanical models

In order to model simultaneous propagation of fractures, a computer program developed by Olson[10] that is based on the conceptual formulation of joint growth[11] was utilized. This methodology utilizes a failure criterion and a propagation velocity model[12] given by:

$$v = A\left(\frac{K_I}{K_{IC}}\right)^n \tag{3}$$

where K_{IC} is the critical fracture toughness and n is subcritical index. The fractures in this methodology are represented by series of equal-length boundary elements. Fracture pattern development is strongly influenced by the mechanical interactions of fractures through the fracture growth history. Based on the mechanical interaction behavior of nearby cracks and effects of other geological information, a fracture length model for larger opening mode fractures propagating through a material with randomly distributed, parallel flaws can be developed. The model requires input geological information such as reservoir thickness, subcritical index, size of stress field, stress boundary conditions and rock properties etc. The boundary element code assumes vertical fractures that are layer bound.

The fracture patterns shown in Figure 5 are generated using the geomechanical model and correspond to variations in the strain value $\left(10^{-3}m, 10^{-4}m, 10^{-5}m\right)$. The sub-critical index value is held constant at $\left(60\,Mpa.\sqrt{m}\right)$ and so is the bed thickness (10m). It is evident that the fracture patterns can be quite different corresponding to different geological conditions. With an increase in stress displacement, the number of fractures in the system increases and the pattern complexity also increases. Similar numerical experiments can be performed by varying the sub-critical index and bed thickness values. While physically realistic fracture patterns can be generated using the geomechanical model, currently the volume of investigation using such models is restricted to small areas of the stress field adjacent to flaws. The cost of simulation will increase significantly if the model is extended to a reservoir scale.

5. Incorporating information from geomechanical model

5.1 Simulation Approach

A key issue that remains to be addressed is the integration of pattern information from analog models together with information from geomechanical models so as to develop a stochastic model for the spatial distribution of subsurface fractures that is physically realistic as well as permits assessment of uncertainty. Fracture pattern information can normally be obtained from well logs or outcrop. However, since only indirect inference of the stress field is possible using borehole image and well core data, there is uncertainty in the predicted stress conditions and that has to be quantified. This uncertainty in reservoir stress values adds to the uncertainty in pattern information inferred on the basis of geomechanical simulations and has to be rigorously accounted for in the multiple-point geostatistical simulation technique.

The uncertainty in reservoir stress condition corresponding to an observed fracture pattern in well logs can be calibrated by applying Bayes' Theorem. Supposing T_{obs} is the fracture pattern observed in a borehole image. Using the geomechanical model and assuming a range of boundary stress values, fracture patterns corresponding to each boundary stress value can be simulated. Corresponding to each stress value B_i, a suite of fracture models can be generated by randomly locating the initial flaw locations. Other geomechanical parameters such as sub-critical index and layer thickness are measured independently and are assumed to be reliably known. These parameters are held constant during the geomechanical simulations. The probability of the fracture pattern T_{obs} in the K models corresponding to a particular boundary stress value B_i can be retrieved. The procedure is repeated for the N boundary stress values $B_i, i=1,..,N$. At the end of this step, the conditional probability $\text{Prob}\{\text{T}_{\text{obs}} \,|\, B_i, i=1,..,N\}$ is obtained.

The likelihood of boundary stress value given an observed fracture pattern - $\text{Prob}\ \{B_i \,|\, \text{T}_{\text{obs}}\}$ can be calculated using Bayes' Theorem:

$$\text{Prob}\ \{B_i \,|\, \text{T}_{\text{obs}}\} = \frac{\text{Prob}\ \{\text{T}_{\text{obs}} \,|\, B_i\} \cdot \text{Prob}\ \{\text{B}_\text{i}\}}{\text{Prob}\ \{\text{T}_{\text{obs}}\}} \tag{4}$$

The $\text{Prob}\{\text{T}_{\text{obs}} \,|\, B_i, i=1,..,N\}$ have been calibrated using the procedure outlined earlier. $\text{Prob}\ \{B_i\}$ is the prior probability corresponding to the stress value B_i. In the absence of any expert information, we can assume each stress value to have the same prior probability i.e. $\text{Prob}\{B_i\} = \frac{1}{N}$. The probability $\text{Prob}\{\text{T}_{\text{obs}}\}$ is obtained concurrently with $\text{Prob}\{\text{T}_{\text{obs}} \,|\, B_i, i=1,..,N\}$ and is equal to:

$$\text{Prob }\{\text{T}_{\text{obs}}\} = \sum_{i=1}^{N} \text{Prob }\{\text{T}_{\text{obs}} \mid B_i\} \cdot \text{Prob }\{\text{B}_{\text{i}}\} \tag{5}$$

i.e. it is the probability of observing the pattern T_{obs} over the entire suite of $N \cdot K$ geomechanical fracture models.

The application of Expression (4) yields the updated distribution for the boundary stress values. This updated probability distribution is denoted as $\text{Prob*}\{B_i\}$. In the stochastic simulation phase, at any step corresponding to a template partially filled with conditioning data (original data plus previously simulated values), the K images corresponding to a particular boundary stress value B_i are scanned for obtaining the probability of fracture patterns in the remaining empty nodes of the spatial template. This yields the probability $\text{Prob}\{\text{A}_{\text{o}} \mid t(n), B_i\}$ where A_0 implies the simulation data event, $t(n)$ is the partially filled fracture pattern. This probability is multiplied by the updated probability $\text{Prob*}\{B_i\}$ to obtain the posterior probability corresponding to the simulation data event $\text{Prob*}\{\text{A}_{\text{o}} \mid t(n), B_i\}$. By repeating this for all boundary stress values, the complete posterior CCDF characterizing the remainder uncertainty in fracture pattern can be constructed. The fracture pattern is propagated by sampling randomly from this posterior CCDF.

Fracture patterns simulated in this fashion rigorously incorporate the uncertainty in fracture pattern characteristics due to the lack of complete knowledge about the underlying physical process for fracture propagation. In addition, the models also incorporate the uncertainty in boundary stress values. Since the calibration process commences from the fracture patterns observed in image logs, the outlined approach is a viable technique for incorporating well log information into stochastic models for the fractured reservoir.

5.2 Discussion

Figure 6 is a fracture pattern observed on a log image. Figure 7 shows the fracture pattern corresponding to a stress displacement value of $8 \cdot 10^{-4}\ m$ and corresponding to two different initial distributions of Griffith cracks. Seven different stress displacement values were assumed and six different fracture patterns corresponding to each stress value were generated by varying the initial flaw locations randomly. As discussed earlier, the prior distribution of the stress values is assumed to be uniform (maximum uncertainty). Figure 8 shows the updated probability distribution of stress values based on the observed fracture pattern depicted in Figure 7. The posterior distribution indicates that the likelihood of the reservoir stress value being of the order $1 \cdot 10^{-4}\ m$ is higher. Better discrimination of the stress displacement value is possible if the pattern T_{obs} retrieved from well logs is more specific. In this case a generic pattern was retained for demonstration purposes.

Figure 9 is the final fracture pattern incorporating the uncertainty in reservoir stress conditions and variations in fracture pattern characteristics observed in the geomechanical models. We can observe that the simulation model has combined the fracture characteristics observed in the suite of geomechanical models such as the NW-SE orientation fractures, the occasional horizontal fractures observed in the some geomechanics models that have no obvious vertical fractures. Some other geomechanical models exhibit short vertical fractures that are also represented in the final simulation model. Another important characteristic of the geomechanics model is that some fractures propagate and terminate against previously existing fractures. This is physically plausible since pre-exiting fractures may reduce the stress at the tip of the daughter fractures, thereby causing the fracture propagation to stall. These characteristics can also be observed with some short fractures terminating against other fractures in simulation model.

The accuracy and robustness of the simulated fracture model is dependent upon the characteristics of the fracture pattern interpreted from image logs. If that pattern is highly specific, the resolution of the stress conditions will be more specific and consequently, only the dominant fracture patterns corresponding to that stress value will be manifested in the final simulation image. Nevertheless, it is possible to generate realistic fracture patterns using the proposed methodology to synthesize information from geomechanical model and well logs.

6. Conclusions

The research focused on developing a methodology for generating physically realistic models of fracture systems in reservoirs. The methodology hinges on the availability of training models of analogous fracture systems. When modeling a target reservoir, the multiple point statistical measures characterizing the patterns observed in the analog can be imposed on the model using a growth-based stochastic simulation technique proposed in this research.

Fracture initiation and growth are affected by a variety of physical geomechanical factors such as the regional stress field, spatial variations of rock properties, or bed thickness. The final model of the reservoir has to integrate the information obtained from geomechanical models and from analog outcrops; in order to yield more physically realistic representation of fracture systems. Furthermore, since important parameters such as the reservoir stress conditions can be only indirectly inferred, the uncertainty in stress field should be quantified and incorporated into stochastic models of the reservoir. That uncertainty can be rigorously quantified using the Bayesian procedure outlined in this paper. The Bayesian procedure is used to update a prior model for uncertainty in reservoir stress field into a posterior model based on the observed image log pattern. This updated probability of reservoir stress values is used to guide the selection of fracture growth patterns during the stochastic simulation phase of the model. Preliminary results obtained using the proposed procedures appear promising.

References

Nelson. Ronald A., 1985, "*Geologic Analysis of naturally fractured Reservoir*", Gulf Publishing Company, Houston, Texas.

Handin, J. and Hager, R.V., 1957, "Experimental determination of sedimentary rocks under confining pressure: Test at room temperature in dry samples", AAPG Bulletin, Vol. 41. pp 1-50.

Stearns, D.W. and Friedman, M., 1972, "*Reservoirs in Fractured Rock*", AAPG Memoir 16, pp 82-100.

Liu, X., Srinivasan, S. and Wong, D.W., 2002, "Geological characterization of naturally fractured reservoirs using multiple point geostatistics", SPE 75246, SPE/DOE Symposium on Improved Oil Recovery.

Atkinson, B.K., and Meredith, P.G., 1987, "The theory of subcritical crack growth with applications to minerals and rocks", In: Fracture mechanics of Rock (edited by Atkinson, B.K.). Academic Press London, 111-166.

Baecher, G.B., Einstein, H.H. and Lanney, N.A, 1977, "Statistical Descriptions of Rock Properties and Sampling", Proc. Of the 18th U.S. Symposium on Rock Mechanics, pp. 5C1.1-5C1.8.

Deutsch, C.V. and Journel, A.G., 1992, *GSLIB; Geostatistical Software Library and User's Guide*, Oxford University Press, New York, N.Y.

Guardiano, F. and Srivastava, R.M., 1992, " Multivariate Geostatistics: Beyond bivariate Moments", in Geostatistics Troia 92, A. Soares editor, Kluwer Acdemic Publisher, Vol. 1, pp. 133-144.

Journel, A.G., 1992, "Geostatistics, Roadblocks and Challenges", in Geostatistics Troia 92. A. Soares editor, Kluwer Academic Publisher, Vol. 1, pp. 213-224.

Olson, J. E., Holder, J. and Rijken, M.C., 2002, "Quantifying the fracture mechanics properties of rock for fractured reservoir characterization", SPE/ISRM 78207.

Segall, P., 1984, "Formation and Growth of Extensional Fracture Sets", Geological Society of America Bulletin 95, 454-462.

Atkinson, B.K., 1984, "Subcritical Crack Growth in Geological Materials*",* Journal of Geophysical Research 89, 4077-4114.

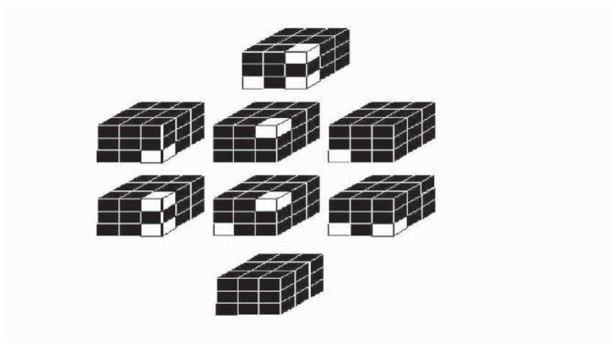


Figure 1: A 27-point 3D spatial template with 24 nodes identified.

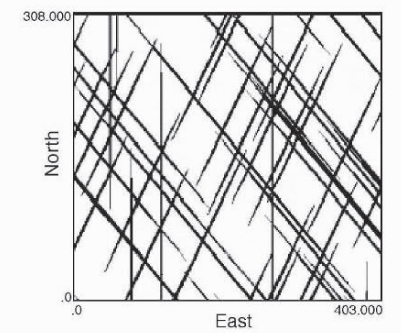


Figure 2: Training image generated object-based simulation.

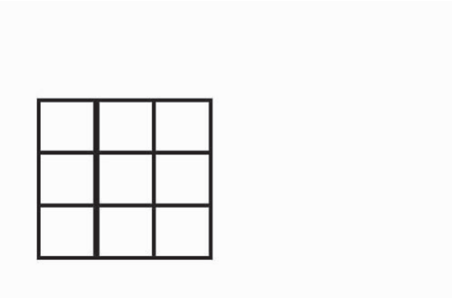


Figure 3.
conditional probability distribution from analog model

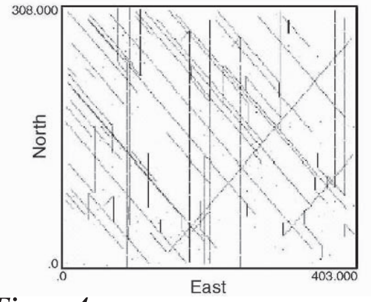
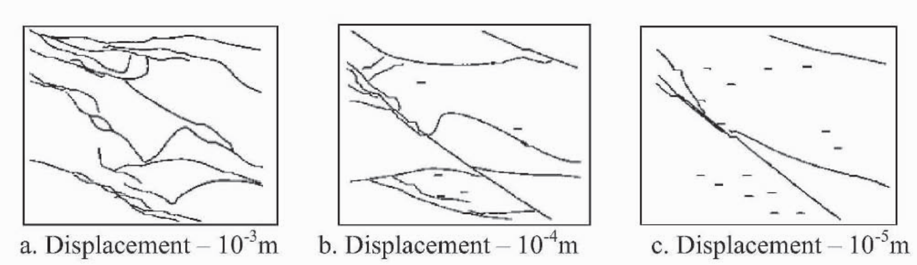


Figure 4.
fracture growth based on training model



a. Displacement – 10^{-3}m b. Displacement – 10^{-4}m c. Displacement – 10^{-5}m

Figure 5. Fracture patterns corresponding to different strain value given constant subcritical index and bed thickness

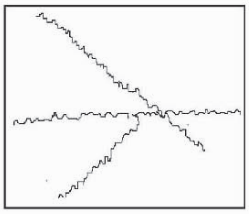


Figure 6. Analog image corresponding to displacement 10^{-4}m.

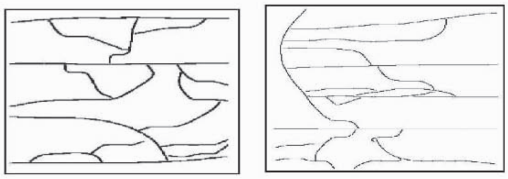


Figure 7. Patterns generated by geomechanics model with the displacement 8x10^{-4}m

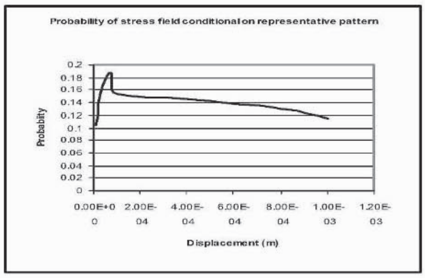


Figure 8. Probability of stress field conditional on observed pattern

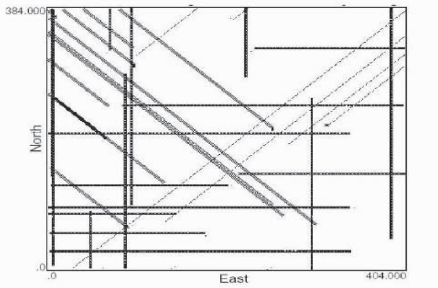


Figure 9. Final simulation image with uncertainty integration in stress field

DIRECT GEOSTATISTICAL SIMULATION ON UNSTRUCTURED GRIDS

JOHN MANCHUK, OY LEUANGTHONG and CLAYTON V. DEUTSCH
Department of Civil & Environmental Engineering, 220 CEB,
University of Alberta, Canada, T6G 2G7

Abstract. Unstructured grids are commonly used in reservoir modeling and are being increasingly considered in complex mining engineering applications. Block kriging of the attributes can be easily implemented; however, this implicitly assumes linear averaging, which is not the case after Gaussian transformation or with variables such as permeability. Direct simulation has been proposed as a solution; however, there are a number of important implementation considerations. This paper addresses the following considerations: (1) search for nearby relevant block and point data, (2) stabilization of the kriging equations and weights in presence of complex screening, (3) correction of the homoscedastic kriging variance to account for realistic proportional effect, (4) determination of valid conditional distribution shapes, (5) accounting for geological controls including stratigraphic surfaces and mixture of multiple facies within an unstructured grid block, and (6) accounting for directional permeability that does not average linearly. Direct simulation on unstructured grids is made practical by addressing these six considerations.

1 Introduction

Unstructured grids are used to model the complex geology and geometry of reservoirs and to provide better accuracy to important development areas. For example, tartan grids are used to provide a high cell density near wells and low cell density in less influential areas (Tran, 1995).

Sequential Gaussian simulation (SGS) (Isaaks, 1990) has become the most extensively used algorithm for continuous variable simulation; however, it is impractical when considering multiscale data, particularly when the data do not average linearly. Direct sequential simulation (DSS) (Xu and Journel, 1994) is an attractive alternative due to the increasing popularity of unstructured grids and the need to integrate multiscale data.

One advantage of DSS is that a wide variety of volume supports can be integrated. This requires that kriging is based on mean covariance/variogram values. There are various ways in which mean covariance calculations can be made more efficient (Pyrcz and Deutsch, 2002). While computational efficiency in this regard is important, an efficient search for nearby relevant data is just as important for practical implementation. The popular method when dealing with regular grids is the super block search strategy, a variation of which could be applied to unstructured grids; however, it may be advantageous to consider different search tree algorithms that may be more efficient.

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 85-94.

The effects of screening remain an issue in the case of multiscale data. Proper filtering of data prior to kriging may be required to avoid anomalously high weights that may lead to extreme estimates. Some filtering techniques such as the octant search, iterative kriging, and the template technique have been used to mitigate screening.

The use of simple kriging (SK) results in an estimation variance that is independent of the data values; this independence is referred to as homoscedasticity. Unfortunately, real data may exhibit a heteroscedastic feature known as the proportional effect, wherein the local mean and variance are often quadratically related (Journel and Huijbreghts, 1978). This heteroscedasticity must be accounted for. An advantage of SK is that covariance reproduction only requires that the mean and variance of this distribution be defined by the SK mean and variance (Journel, 1994). A method of determining the local distribution shapes has been developed and will be revisited (Pyrcz and Deutsch, 2002; Deutsch et al, 2001; Oz et al, 2001).

The advantage of unstructured grids in capturing more complex geology also entails further complications related to geological controls such as stratigraphic surfaces and a mixture of multiple facies that may be represented within any particular block. An unstructured grid may not conform to the stratigraphic setting, which introduces problems relating to selecting relevant data for kriging and estimating grid blocks that contain multiple subsequence layers.

Further, the use of average variogram/covariance values in SK (inside DSS) for multiscale data has an implicit assumption of linear averaging of the model variables. This poses a problem when the variable of interest does not average linearly. Permeability is a classic example of such a variable. Accounting for the appropriate type of averaging is integral to the correct implementation of DSS.

This paper addresses these six important issues and proposes some novel approaches for resolution.

2 Search for Nearby Relevant Block and Point Data

When considering unstructured grids, data may consist of original data at a small scale, regularly gridded soft data, and grid blocks of varying sizes. There are several methods that can be used to deal with this array of data: A brute force method involving a matrix of distances n_{GB} by n_{GB} in size, where n_{GB} is the number of grid blocks; A super block search strategy (Deutsch and Journel, 1998); or the use of search trees. The brute force method is only applicable to small problems as larger data sets would be impractical for conventional computer memory availability. A super block strategy could be used; however, implementing certain types of search trees will be more efficient.

One type of search trees common in the computer graphics and computer gaming industry are quadtrees and octrees, (Figure 1) (Frisken and Perry, 2002). When considering graphics visualization, these search trees are used to quickly determine which polygons are in view such that only those polygons are drawn (this reduces memory requirements). Searching for nearest neighboring data is a similar task.

Quadtrees and octrees organize data in such a way that point location, region location, and nearest neighbor operations can be done easily. Frisken and Perry (2002) introduce a binary indexing system for quadtrees that allows for efficient execution of the above operations. This system can easily be applied to octrees for three dimensional data as well.

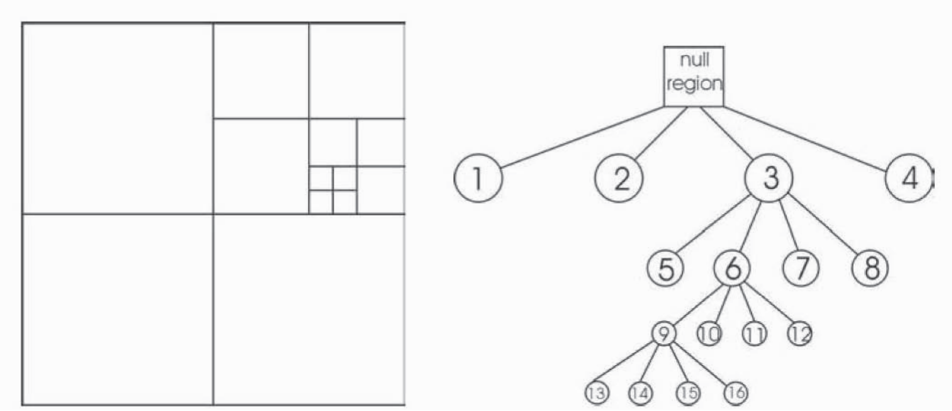


Figure 1: Example of a quadtree structure (left) and tree-representation (right). The quadtree is more refined in areas with higher data densities.

Implementing quadtrees or octrees to organize spatial data for simulation purposes allows for efficient acquisition of nearby data for each node to be simulated. The search for nearest neighboring leaf nodes is not dependant on the type of grid and tree traversal for finding and inserting points is a simple process. Having these characteristics along with low memory requirements makes search trees excellent for unstructured grid problems.

3 Stabilization of the Kriging Equations and Weights in the Presence of Complex Screening

Screening can cause extreme positive and negative weights that lead to erroneous estimates and estimation variances. One method of reducing the occurrence of extreme weights is to remove data from the kriging matrix: this iterative kriging technique will remove data until the absolute value of all the weights are below a specified maximum. Iterative kriging works; however, data that may be highly influential in estimating a location could be removed from the kriging matrix resulting in a less accurate result. Another method of reducing screening is the template technique which involves rejecting any data that are shadowed by a closer data (Figure 2). A downfall to the template technique is its high demand on computation time.

A new method of filtering data used in estimating a location is the sector search method, which is somewhat similar to the template technique. The sector search method uses input dip and azimuth tolerances to create sectors in which only the nearest data is selected for kriging (Figure 2).

The sector search subroutine works fast in two dimensions as the sectors are all pre-constructed and then translated to locations of interest; however, in three dimensions, the sectors are built as points are encountered making the process more time consuming.

Even though the sector search method removes many screened data, there may be unreasonable screening still present. For example, consider two points in adjacent sectors, the point closer to the location being estimated will screen the effect of the second point. Using larger sectors will keep screening to a minimum.

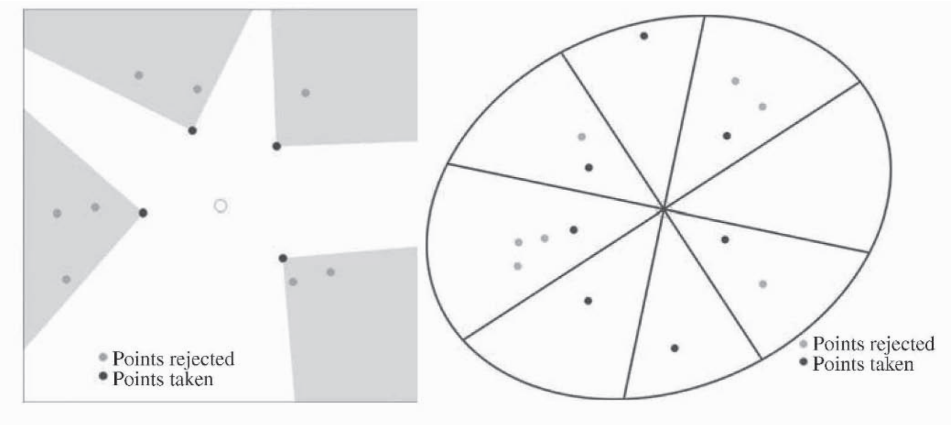


Figure 2: The template technique (left) and the sector search method (right) to reduce screening.

4 Correction of the homoscedastic kriging variance to account for realistic proportional effect

Data in original units are often heteroscedastic. High valued areas are more variable. This heteroscedastic behavior is commonly referred to as the proportional effect (Journel and Huijbregts, 1978). DSS relies on covariance reproduction through local distributions whose mean and variance are defined by SK (Xu and Journel, 1994). For data following the congenial Gaussian distribution this assumption is correct; however for data exhibiting heteroscedastic features, it is an unsuitable assumption due to the homoscedasticity of the kriging variance. The kriging variance must be adjusted such that the proportional effect is reproduced.

4.1 DSS USING LOGNORMAL DATA

To see the effects of directly simulating data that exhibit the proportional effect, a study was performed using a lognormal distribution. This distribution was chosen for a number of reasons: (1) although real data are not necessarily lognormal, most data exhibit a strong asymmetry similar to that characterized by the lognormal distribution, and (2) there is a clear mathematical link between the lognormal and the more common

Gaussian distribution that permits tractability of the results. Further, an equation describing the proportional effect of lognormal data exists (Journel and Huijbregts, 1978). Knowing these relations, the kriging variance can be calibrated to honor the heteroscedasticity inherent in lognormal data.

An exhaustive lognormal data set was generated by transforming an unconditional Gaussian model (Figure 3). The mean and variance of the lognormal data were arbitrarily chosen to be 100 and 10000, respectively. A set of 625 samples was drawn from the model and used for numerical experimentation.

4.1.1 *Options of Simulation Explored*

Three options were identified for evaluation:

Option 1	Perform SGS
Option 2	Perform DSS without correcting the kriging variance
Option 3	Perform DSS and correct the kriging variance to honor the proportional effect

With lognormal data, an equation exists for correcting the variance using the mean or estimate:

$$\sigma^2_{Z,C} = \left[z^*(\mathbf{u})\right]^2 (e^{\beta^2_G \cdot \sigma^2_Y} - 1)$$

Where $\sigma^2_{Z,C}$ is the corrected variance, σ^2_Y is the local variance in normal space, and β^2_G is the global variance of *ln(Z)*. By determining a relation between the estimation variance in Gaussian space and that in lognormal space, the value of σ^2_Y could be determined without having to perform kriging twice.

For each option, 100 realizations were generated and the E-type mean and variance was calculated (Figure 4). Reproduction of the global statistics and the variogram were verified. Figure 4 also shows similar results between DSS with a correction and SGS; however, with DSS and no correction as in Option 2, the variance is clearly homoscedastic. A more visual comparison of the three options is available in Figure 5 where the spatial distribution of the mean and standard deviation show reproduction of the proportional effect in a single realization.

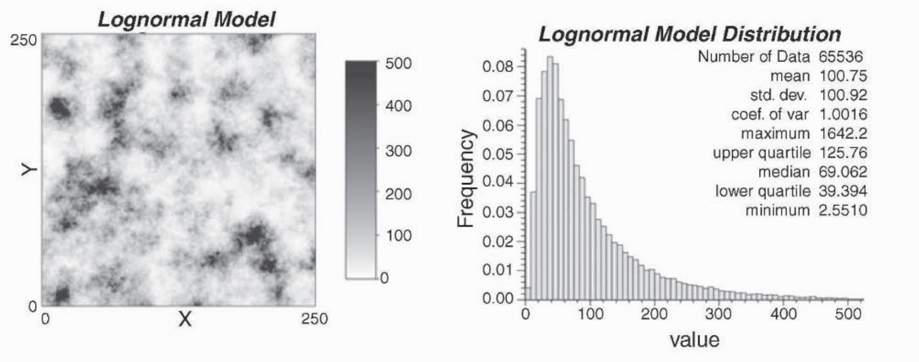


Figure 3: The lognormal model used for direct simulation experimentation.

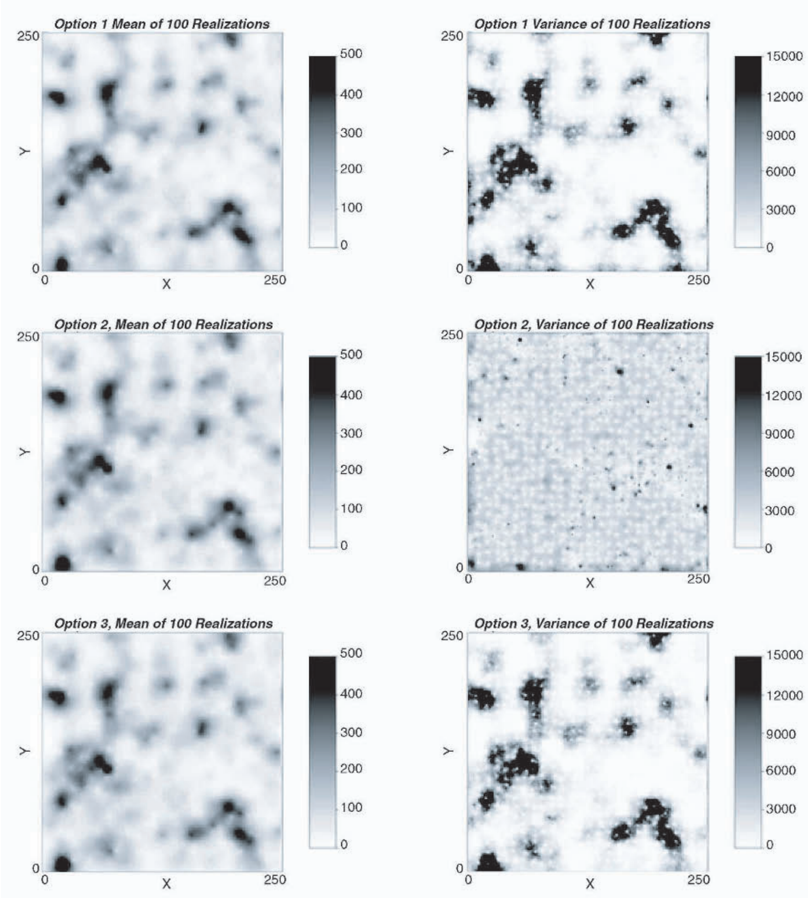

Option 1 Mean of 100 Realizations
Option 1 Variance of 100 Realizations
Option 2, Mean of 100 Realizations
Option 2, Variance of 100 Realizations
Option 3, Mean of 100 Realizations
Option 3, Variance of 100 Realizations
X
Y
0
250
500
400
300
200
100
15000
12000
9000
6000
3000

Figure 4: The mean and variance taken over 100 realizations for all three simulation approaches: Option 1 is the straightforward SGS, Option 2 refers to DSS, and Option 3 refers to DSS with variance correction to account for heteroscedasticity.

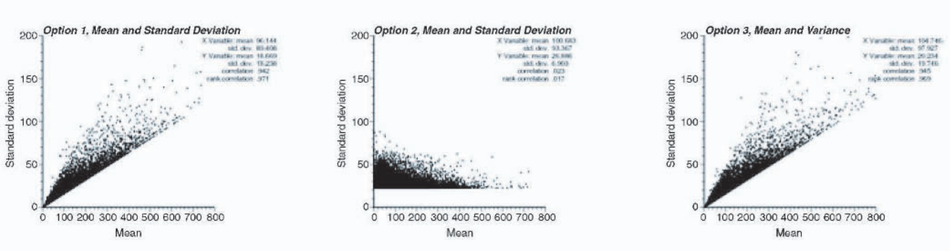


Figure 5: Local mean versus standard deviation at every estimated location for options 1 (left), 2 (middle), and 3 (right). Options 1 and 3 show the proportional effect and compare nicely. Option 2 shows a homoscedastic variance.

By performing simulation using lognormal data, it is possible to introduce a solution for dealing with the proportional effect. The lognormal distribution is particularly useful because the proportional effect is one of its prominent features and is analytically accessible.

We expect that the proportional effect could be fit from real data instead of using either the Gaussian model of no proportional effect of the lognormal model of a quadratic proportional effect.

5 Determination of valid conditional distribution shapes

A method to determine the shape of the local distributions in original units from the SK mean and variance is needed such that the global distribution is reproduced.

Figure 6 shows a numerical integration approach proposed by Oz et. al. (2001). If a specific probability *p* of a non-standard normal distribution with mean *m* and standard deviation *σ* is known, the corresponding direct space quantile can be calculated as:

$$Z(\mathbf{u}) = F^{-1}[G_{\{0,1\}}[G^{-1}{}_{\{m,\sigma\}}(p)]]$$

where $G_{\{0,1\}}$ is the cumulative distribution function (cdf) of a standard Gaussian distribution, $G_{\{m,\sigma\}}$ is the cdf of a non-standard Gaussian distribution with mean *m* and standard deviation *σ*, and F is the cdf of the representative data distribution. Z(**u**) is the *p* quantile of the local distribution of uncertainty (Pyrcz and Deutsch, 2002).

By creating a series of *Y*-space distributions from a list of means and variances and repeating the above procedure for a range of quantiles, a set of *Z*-space distributions can be generated. The mean and variance of each *Z*-space distribution can be calculated and used as reference values. Upon kriging at a particular location, the resulting mean and variance can be used to look up the corresponding local distribution in original units, from which a simulated value can be drawn.

6 Accounting for geological controls including stratigraphic surfaces and mixture of multiple facies within an unstructured grid block

Some geological settings are characterized by a series of genetically related strata. The geology may consist of a sequence stratigraphic framework; the bounding surfaces between the layers correspond to a specific geologic time that separates two different periods of deposition or a period of erosion followed by deposition (Deutsch, 2002). This presents some potential issues related to the structure such as: the grid does not line up with the stratigraphic surfaces, grid blocks may contain multiple facies and subsequences (Figure 7), and searching for relevant data to estimate unknown locations.

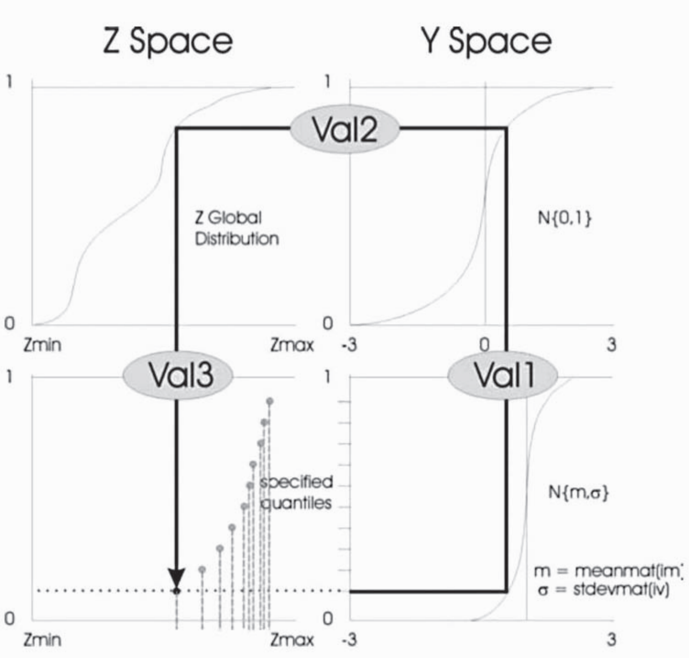


Figure 6: The graphical representation of the transformations applied to calculate the local distributions of uncertainty with a shape such that the global distribution is reproduced. The illustrated transformation is repeated for a sufficient number of quantiles to describe the local distribution.

A possible method of dealing with data within various subsequences is to flag the data by subsequence and only use data within genetically related strata. When simulating blocks that cross multiple subsequences, flagging and simulating its value poses a problem. One idea is to discretize the block into smaller “blocks”, flag the smaller components and estimate them to obtain a value or multiple values and structure within a grid block. Since blocks may cross into multiple subsequences as well as contain multiple facies, a method to determine that portion of a grid block relevant to estimation is required. An idea of the subsequence structure within grid blocks being estimated as well as those being used for conditioning data is critical (Figure 8).

Upon estimating grid blocks, the proportion of facies within each block can be determined overall, but it may be better to retain the facies proportions within each subsequence in a block.

7 Accounting for directional permeability that does not average linearly

Because data exist in vastly different scales such as small core-based permeability and large scale production data, problems arise due to the scale difference and non linear averaging of permeability. By implementing a power law transform, permeability values will approximately average linearly, and can then be used in a direct simulation approach (Zanon et al, 2002).

The general formulae for power law averaging is

$$K_{eff} = \left[\frac{1}{v} \int_{v} k(\mathbf{u})^{\omega} d\mathbf{u} \right]^{\frac{1}{\omega}}$$

where v is the volume over which the average is calculated, $k(\mathbf{u})$ is the permeability at location $\mathbf{u}$ in v, and ω is an averaging exponent.

Since DSS utilizes kriging as an estimator, the model variables must average linearly with scale. By using a power law transformation prior to kriging, the problems generated by multiscale data can be avoided and transformed variables will average linearly with scale. A Gaussian transform would undo the benefit of the power-law transform; it is important to perform kriging and simulation in the correct units.

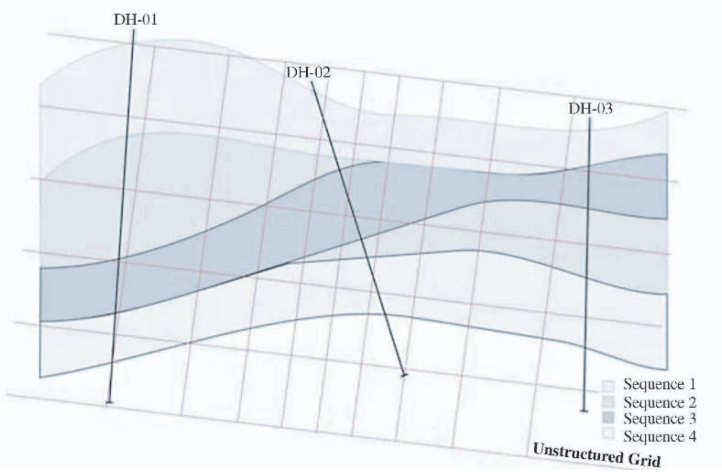


Figure 7: Stratigraphic surfaces and superimposed unstructured grid. Three hypothetical drill holes/wells are also shown.

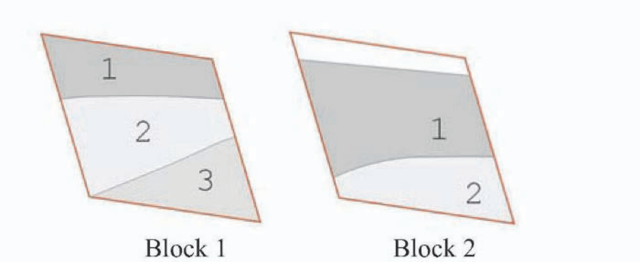


Figure 8: Unstructured grid block crossing multiple subsequence layers. If block 1 is being estimated using block 2, only data within block 2 and subsequence 1 should be used to estimate data within block 1 subsequence 1.

A concern of implementing power law transformation, especially when dealing with unstructured grids, is that ω may not be constant over every volume support. An unstructured grid may involve many different volume support sizes to be estimated and when the scale difference is large, ω may change. Other concerns that affect the value of ω are arbitrarily chosen boundary conditions and if the formation approaches the percolation threshold (Kirkpatrick, 1973).

8 Conclusions

Unstructured grids are practically relevant for realistic reservoir modeling. The distinction of simulating in the units of the original data provides significant benefits such accounting for multiscale data and permitting different local distributional shapes. In practice, implementation of DSS has been limited. Even something as seemingly straightforward as searching for data is complicated by the multiscale nature of the problem. In these instances, quadtrees or octrees may be particularly efficient. Screening may also lead to destabilization of the kriging matrix, thus a preferential filtering of the data through a sector search may be appropriate. Multiscale issues are further complicated by the very nature of the model variable, whether these variables average linearly or whether pre-processing transform such as the power law transform is required.

Unstructured grids allow for increasingly complex geology to be integrated; however, this presents issues in grid block definition and facies identification if the blocks are too large and/or if they cross multiple sequence or sub-sequence stratigraphic layers. Despite all these issues, perhaps the most important advance presented in this paper is the correction applied to the SK variance to account for the heteroscedastic nature that is often inherent to real data. The lognormal case was used to illustrate a corrective approach to effectively reproduce heteroscedasticity.

References

Deutsch, C.V., *Kriging with Strings of Data*, Mathematical Geology, Vol. 26, No. 5, 1994.

Deutsch, C.V., *Geostatistical Reservoir Modeling*, Oxford University Press Inc., 2002.

Deutsch, C.V. and Journel, A.G., *GSLIB Geostatistical Software Library and User's Guide*, Second Edition, Oxford University Press, 1998.

Deutsch, C.V. Tran, T. and Xie, Y., *A preliminary report on: An approach to ensure histogram reproduction in direct sequential simulation,* Technical report, Center for Computational Geostatistics, University of Alberta, Edmonton, AB, March 2001.

Frisken, S.F., and Perry, R.N., *Simple and Efficient Traversal Methods for Quadtrees and Octrees*, Mitsubishi Electrical Research Laboratories, 2002.

Journel, A.G. and Huijbregts, Ch.J., Mining Geostatistics, Academic Press Limited, 1978.

Kirkpatrick, S., *Percolation and Conduction*, Reviews of Modern Physics, 1973, Vol. 45, No. 4, pp 574-588

Leuangthong, O. and Deutsch, C.V., *Modeling Multivariate Multiscale Data*, Center for Computational Geostatistics, University of Alberta, Report 4: 2002.

Oz, B., Deutsch, C.V., Tran, T., and Xie, Y., *A fortan 90 program for direct sequential simulation with histogram reproduction*, Computers & Geosciences, page submitted 2001.

Tran, T.T., *Stochastic Simulation of Permeability Fields and Their Scale-up for Flow Modeling*, PhD Thesis, Stanford University, 1995.

Xu, W. and Journel, A.G., *DSSIM: A General Sequential Simulation Algorithm,* Stanford Center for Reservoir Forecasting, May 1994.

Zanon, S. Nguyen, H. and Deutsch, C.V., Power Law Averaging Revisited, Center for Computational Geostatistics, University of Alberta, Report 4, 2002.

DIRECTIONAL METROPOLIS: HASTINGS UPDATES FOR POSTERIORS WITH NONLINEAR LIKELIHOODS

HÅKON TJELMELAND
Department of Mathematical Sciences, NTNU, 7491 Trondheim, Norway
JO EIDSVIK
Statoil Research Center, 7005 Trondheim, Norway

Abstract. In this paper we consider spatial problems modeled by a Gaussian random field prior and a nonlinear likelihood linking the hidden variables to the data. We define a directional block Metropolis–Hastings algorithm to explore the posterior. The method is applied to seismic data from the North Sea. Based on our results we believe it is important to assess the actual posterior in order to understand possible shortcomings of linear approximations.

1 Introduction

Several applications in the earth sciences are preferably formulated by an underlying hidden variable which is indirectly observed via noisy measurements. Examples include seismic data, production data and well data in petroleum exploration: In seismic data the amplitudes are nonlinearly connected to the elastic parameters of the subsurface, see e.g. Sheriff and Geldart (1995). Production data contain the history of produced oil and gas, which is a complex functional of the permeability properties in the reservoir, see e.g. Hegstad and Omre (2001). Well data of radioactivity counts need to be transformed into more useful information, such as clay content in the rocks, see e.g. Bassiouni (1994). The Bayesian framework is a natural approach to infer the hidden variable; this entails a prior model for the variables of interest and a likelihood function tying these variables to observations.

In this paper we consider Gaussian priors for the underlying spatial variable, and nonlinear likelihood models. When using a nonlinear likelihood, the posterior is not analytically available. However, the posterior can be explored by Markov chain Monte Carlo sampling (see e.g. Robert and Casella (1999)), with the Metropolis–Hastings (MH) algorithm as a special case. We describe a directional MH algorithm in this paper, see Eidsvik and Tjelmeland (2003). We use this algorithm to update blocks of the spatial variable at each MH iteration. We show results of our modeling procedures for seismic data from a North Sea petroleum reservoir.

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 95-104.

2 Methods

2.1 PRIOR AND LIKELIHOOD ASSUMPTIONS

The variable of interest is denoted $x = \{x_{ij} \in \mathcal{R}; i = 1, \ldots, n; j = 1, \ldots, m\}$, a spatial random field in two dimensions represented on a grid of size $n \times m$. We denote its probability density by $\pi(x) = \pi(x|\theta) = N(x; \mu(\theta), \Sigma(\theta))$, where $N(x; \mu, \Sigma)$ denotes a Gaussian density evaluated in x, with fixed mean μ and covariance matrix Σ. For generality we condition on hyperparameters θ, but in this study we treat θ as fixed parameters. The generalization to a vector variable, $x_{ij} \in \mathcal{R}^d$, is straightforward. For the application in Section 3 we have $x_{ij} \in \mathcal{R}^3$. A three dimensional grid, x_{ijk}, is of course also possible.

We assume the spatial variable x to be stationary and let the field be defined on a torus, see e.g. Cressie (1991). As explained below, this has important computational advantages, but the torus assumption also implies that one should not trust the results close to the boundary of the grid. Thus, one should let the grid cover a somewhat larger area than what is of interest.

The likelihood model for the data $z = \{z_{ij} \in \mathcal{R}; i = 1, \ldots, n; j = 1, \ldots, m\}$, given the underlying variable x, is represented by the conditional density $\pi(z|x) = N(z; g(x), S)$, where $g(x)$ is a nonlinear function. Hence, the conditional expectation of the data has a nonlinear conditioning to the underlying field. We assume that the likelihood noise is stationary with covariance matrix S. It is again straightforward to extend this model to vector variables at each location, $z_{ij} \in \mathcal{R}^d$, or three dimensional grids. For the application in Section 3 we have $z_{ij} \in \mathcal{R}^2$. We assume that a linearized version of the likelihood is available, and denote this by the conditional density $\pi^{lin}_{x_0}(z|x) = N(z; G_{x_0} x, S)$, where x_0 is the value of x used in the linearization.

The posterior of the hidden variable x conditional on the data is given by

$$\pi(x|z) \propto \pi(x)\pi(z|x), \tag{1}$$

an analytically intractable posterior. The linearized alternative;

$$\pi^{lin}_{x_0}(x|z) \propto \pi(x)\pi^{lin}_{x_0}(z|x), \tag{2}$$

for fixed x_0, can be written in a closed form, and is possible to evaluate and sample from directly. But note that in general this becomes computationally expensive in high dimensions. With our torus assumption discussed above, the covariance matrices involved become circular and the linearized posterior can then be evaluated and sampled from effectively in the Fourier domain (Cressie (1991), Buland, Kolbjørnsen, and Omre (2003)). The actual nonlinear posterior can also be evaluated, up to a normalizing constant, in the Fourier domain, by treating the prior and likelihood terms in equation (1) separately.

2.2 METROPOLIS–HASTINGS BLOCK UPDATING

A MH algorithm is an iterative sampling method for simulating a Markov chain that converges to a desired posterior distribution, see e.g. Robert and Casella (1999). Each iteration of the standard MH algorithm consists of two steps: (i) Propose a new value for the underlying variable, (ii) Accept the new value with a certain probability, else keep the value from the previous iteration.

We describe a MH algorithm which updates blocks of the random field at each iteration. Let $X^i = x$ denote the variable after the i-th iteration of the MH algorithm. For the $(i+1)$-th iteration we draw a block of fixed size $k \times l$ at random, where $k < n$, $l < m$. Since the grid is on a torus, there are no edge problems when generating this block. We denote the block by A, a defined boundary zone of the block by B, and the set of nodes outside the block and boundary by C, see Figure 1. Further, we split the variable into these blocks; $x = (x_A, x_B, x_C)$ as the parts in the block, the boundary, and outside the block and boundary zone, respectively. Correspondingly, we denote the data vector by $z = (z_A, z_B, z_C)$. To propose a new value in the block we define a proposal density for the part in A, and denote the proposal on this block by y_A. The rest of x remains unchanged, and hence the proposed value is $y = (y_A, x_B, x_C)$. One proposal density in the block is the linearized posterior for y_A, conditional only on values in the boundary zone and data in A and B. We denote this by $\pi_x^{lin}(y_A|x_B, z_A, z_B) = N(y_A; m, T)$, where the mean m and covariance T can be calculated directly (Cressie, 1991). Note that the linearization is done at the current value $X^i = x$. The final step in the MH iteration is to accept or reject the proposed value y, and we thus obtain the next state X^{i+1}. Note that the results of the MH algorithm are the same no matter which block size we choose. The CPU time, on the other hand, will vary with the block size. In the application below we have chosen a block size similar to the range of the spatial correlation.

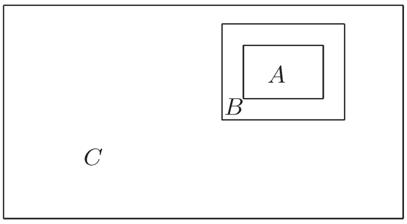


Figure 1. Block updating. The full size of the grid is $n \times m$. We illustrate the block A of gridsize $k \times l$, a boundary zone B, and the other parts of the field C.

2.3 DIRECTIONAL METROPOLIS–HASTINGS

Directional MH algorithms are a special class of MH algorithms. Each iteration consists of the following steps; (i1) Generate an auxiliary random variable which defines a direction, (i2) Draw a proposed value on the line specified by the current state and the auxiliary direction, (ii) Accept the proposed value with a certain probability, else keep the variable from the previous iteration.

We present a directional updating scheme where the proposal step is done effectively in one dimension, see Eidsvik and Tjelmeland (2003). We outline our method following the block sampling setting in Section 2.2. Denote again the variable at the i-th iteration by $X^i = x = (x_A, x_B, x_C)$, and the data by $z = (z_A, z_B, z_C)$. We generate an auxiliary direction (step i1) as follows; First, draw w_A from $\pi_x^{lin}(\cdot|x_B, z_A, z_B)$. Next, define the auxiliary direction as $u = \pm\frac{w_A - x_A}{|w_A - x_A|}$, where we use + or − so that the first component of u is positive, see the discussion in Eidsvik and Tjelmeland (2003). Since this density for w_A is Gaussian, it is possible to calculate the density for the auxiliary unit direction vector u (Pukkila and Rao, 1988). We denote this density by $g(u|x_A, x_B, z_A, z_B)$.

At the last part of the proposal step (i2) we draw a one dimensional value t from some density $q(t|u, x, z)$, and set $y_A = x_A + tu$ as the proposed value for the block, and $y = (y_A, x_B, x_C)$ as the proposal for the entire field. This proposal is accepted, i.e. $X^{i+1} = y$, with probability

$$r(y|x) = \min\left\{1, \frac{\pi(y)}{\pi(x)} \cdot \frac{\pi(z|y)}{\pi(x|y)} \cdot \frac{g(u|y_A, x_B, z_A, z_B)}{g(u|x_A, x_B, z_A, z_B)} \cdot \frac{q(-t|u, y, z)}{q(t|u, x, z)}\right\}, \tag{3}$$

else we have $X^{i+1} = x$.

In particular, if we choose the one dimensional density as

$$q^\star(t|u, x, z) \propto \pi(z|x_A + tu, x_B, x_C)\pi(x_A + tu, x_B, x_C)g(u|x_A + tu, x_B, z_A, z_B), \tag{4}$$

the acceptance probability in equation (3) is equal to unity (Eidsvik and Tjelmeland, 2003), and hence the proposed variable is always accepted. Two elements are important when considering the unit acceptance rate proposal in equation (4); (i) An acceptance rate of one is not necessarily advantageous in MH algorithms. It is advantageous to obtain fast mixing, i.e. to have a small autocorrelation between successive variables. This is best achieved with large moves at each iteration of the MH algorithm. (ii) It is not possible to sample directly from the $q^\star$ density. To obtain a sample we have to fit an approximation to $q^\star$ in some way, either by a parametric curve fit, or by numerical smoothing of a coarse grid approximation.

In this paper we use an alternative density which does not give unit acceptance rate. The adjusted density is

$$\widetilde{q}(t|u, x, z) \propto (1 + |t|)^\lambda q^\star(t|u, x, z), \qquad \lambda > 0, \tag{5}$$

where the $(1+|t|)^\lambda$ term encorages t values away from $t = 0$. This makes it possible to have larger moves in the MH algorithm since $t = 0$ corresponds to $y_A = x_A$. We fit an approximation by first calculating $\widetilde{q}$ on a coarse grid and then use linear interpolation in the log scale between the evaluated grid points. The approximation that we obtain with this approach is our proposal denoted q.

In Figure 2 we show the proposal with acceptance one, $q^{\star}$, the adjusted proposal $\widetilde{q}$ and its fitted proposal q. This particular plot is obtained in our application in Section 3. We tuned λ in equation (5) so that the acceptance rate was about 0.5 in our application. This seems to be a reasonable acceptance rate considering the asymptotically optimal acceptance rates for random walk MH and Langevin MH algorithms at 0.25 and 0.5, respectively, see e.g. Robert and Casella (1999). We obtain this by setting $\lambda = 10$. Note that the proposal density for t has a bimodal shape. One mode is usually close to $t = 0$, while the other mode is quite far from 0. The mixing of the MH algorithm improves if we reach the mode away from 0 more often. This can be established by $\widetilde{q}$ or q as shown in Figure 2. In the case of a linear likelihood the two modes illustrated in Figure 2 are always of equal size. A nonlinear likelihood causes them to have unequal mass, and most commonly the mode near $t = 0$ contains most of the probability mass.

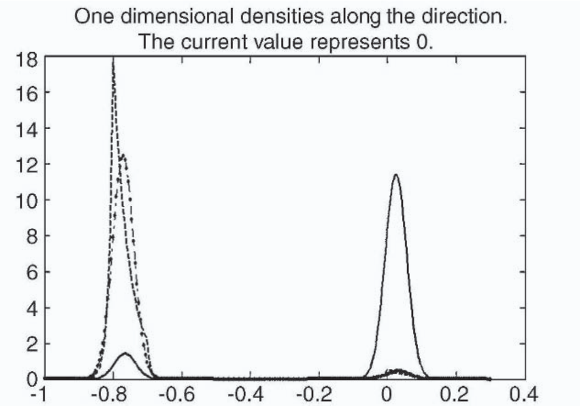


Figure 2. Sketch of the density $q^{\star}$(solid), the adjusted density $\widetilde{q}$ (dash-dots), and the fitted density q (dashed). This is a typical proposal density for t in our application in Section 3. The approximation q has an exponential form in each interval of length 0.1 on a grid (for example between -0.8 and -0.7).

3 Example

3.1 SEISMIC AMPLITUDE VERSUS OFFSET DATA

Seismic amplitude versus offset (AVO) analysis is commonly used to assess the underlying lithologies (rocks and saturations) in a petroleum reservoir, see e.g. Sheriff and Geldart (1995), and Mavko, Mukerji, and Dvorkin (1998). The reflection amplitude of seismic data changes as a function of incidence angle and as a function of the elastic properties which are indicative of lithologies.

We analyze reflection data at two incidence angles in the Glitne field, North Sea. The Glitne field is an oil-producing turbidite reservoir with heterogeneous sand

and shale facies, see Avseth et al (2001) and Eidsvik et al (2004). The domain of our interest is $2.5 \times 2.5\ km^2$, and it is split into a grid of size 100×100, with each grid cell covering $25 \times 25\ m^2$. The area covers what was interpreted as the lobe of the turbidite structure in Avseth et al (2001). Figure 3 shows the reflection amplitude along the grid at reflection angles zero (left) and thirty (right).

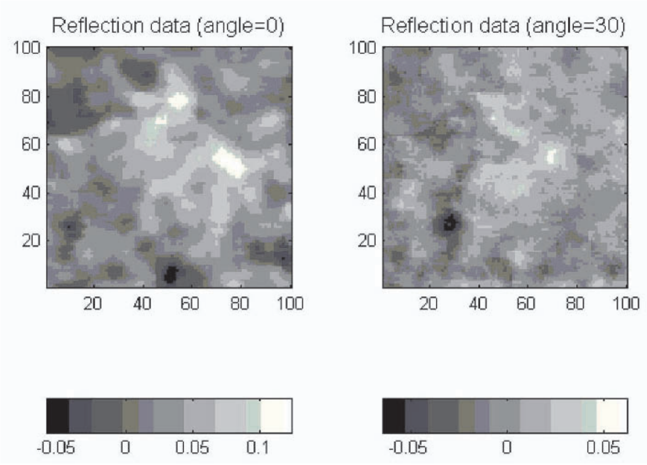


Figure 3. Reflection data at the 2D interface. Incidence angle 0 degrees (left) and 30 degrees (right).

The reflection amplitudes are hard to analyze directly because they are a result of the contrast in elastic properties in the cap rock covering the reservoir and the properties in the reservoir zone. We next use a statistical model for automatic analysis of the elastic reservoir properties.

3.2 STATISTICAL MODEL FOR SEISMIC DATA

The seismic data for the two reflection angles are denoted $z = (z^0, z^1)$, where z^0 refers to the zero offset reflection and z^1 to 30 degrees incidence angle (both are plotted in Figure 3). The statistical model that we use is closely connected to the one in Buland, Kolbjørnsen and Omre (2003).

The variables of interest are the pressure and shear wave velocities, and the density in the reservoir zone. We denote these by $x = (\alpha, \beta, \rho) = \{x_{ij} = (\alpha_{ij}, \beta_{ij}, \rho_{ij}); i = 1, \dots, n; j = 1, \dots, m\}$, where α is the logarithm of the pressure wave velocity, β is the logarithm of the shear wave velocity, and ρ is the logarithm of the density. The velocities and density are the elastic properties of the rocks which in some sense capture the rock mineral and saturation (Mavko, Mukerji, and Dvorkin, 1998). The units for the exponential equivalents of α, β and ρ are m/s for velocities and kg/m^3 for density. Let $(\alpha_0, \beta_0, \rho_0)$ be the logarithm of pressure and shear wave velocities, and the logarithm of density for the cap rock. These cap rock properties are treated as fixed values in this study, and are equal for all locations in the grid.

We assign a Gaussian prior density to the reservoir variables of interest, i.e. $\pi(x) = N(x; \mu, \Sigma)$, where μ now becomes a $3mn$ vector with the mean of the three log elastic properties. These prior mean values are fixed, and set to $\mu_\alpha = E(\alpha_{ij}) = 7.86$, $\mu_\beta = E(\beta_{ij}) = 7.09$, $\mu_\rho = E(\rho_{ij}) = 7.67$, for all (i, j). These mean values are assessed from well logs in Glitne, see Avseth et al (2001). The prior covariance matrix, Σ, is a $3mn \times 3mn$ matrix defined by a Kronecker product, giving a 3×3 block covariance matrix at the diagonal, and 3×3 matrices with this covariance function and a spatial correlation function on the off-diagonal, see Buland, Kolbjørnsen, and Omre (2003). The diagonal covariance matrix describing the marginal variability at each location is defined by $Std(\alpha_{ij}) = 0.06$, $Std(\beta_{ij}) = 0.11$, $Std(\rho_{ij}) = 0.02$, and correlations $Corr(\alpha_{ij}, \beta_{ij}) = 0.6$, $Corr(\alpha_{ij}, \rho_{ij}) = 0.1$, $Corr(\beta_{ij}, \rho_{ij}) = -0.1$. These parameters capture the variability expected from previous studies, see Buland, Kolbjørnsen, and Omre (2003). The spatial correlation function is the same for all three reservoir variables and is an isotropic exponential correlation function with range $250m$ (10 grid nodes).

The likelihood function is nonlinear, and is defined by approximations to the Zoeppritz equations, see e.g. Sheriff and Geldart (1995). The density for the seismic AVO data, given the underlying reservoir properties, is $\pi(z|x) = N(z; g(x), S)$, where the nonlinear function goes only locationwise, i.e. at grid node (i, j) the expectation term in the likelihood is a function of the variables at this gridnode only. For each location (i, j) and angle $\gamma = 0$, 30 we have

$$g_{ij,\gamma}(x) = g_{ij,\gamma}(\alpha_{ij}, \beta_{ij}, \rho_{ij}) = a_0(\alpha_{ij} - \alpha_0) + a_{1,ij}(\beta_{ij} - \beta_0) + a_{2,ij}(\rho_{ij} - \rho_0), \quad (6)$$

where

$$\begin{aligned} a_0 &= \frac{1}{2}[1 + sin^2(\gamma)], \qquad a_{1,ij} = -4\xi_{ij} sin^2(\gamma), \qquad (7) \\ a_{2,ij} &= \frac{1}{2}[1 - 4\xi_{ij} sin^2(\gamma)], \qquad \xi_{ij} = \frac{\exp(2\beta_{ij}) + \exp(2\beta_0)}{\exp(2\alpha_{ij}) + \exp(2\alpha_0)}. \end{aligned}$$

The noise covariance matrix of the likelihood, S, is a $2mn \times 2mn$ matrix defined from a Kronecker product. This covariance matrix has a block diagonal 2×2 matrix on the diagonal, and off-diagonal elements defined from an exponential correlation structure with range $250m$. The diagonal noise covariance matrix is defined by $Std(\gamma = 0) = 0.015$, $Std(\gamma = 30) = 0.012$, $Corr(\gamma = 0, \gamma = 30) = 0.7$. This likelihood noise model is specified using the parameters in Buland, Kolbjørnsen, and Omre (2003).

A linear likelihood model can be defined by fixing the ratio ξ_{ij} in equation (7). For a constant linearization point we have $\pi_\mu^{lin}(z|x) = N(z; G_\mu x, S)$. A similar linearization is used in Buland, Kolbjørnsen, and Omre (2003), and with this linearization they assess the analytically available posterior directly on a $3D$ dataset. For a linearized proposal density on the block A we have $\pi_x^{lin}(\cdot|x_B, z_A, z_B)$, where we use a block size of 9×9, and a boundary zone of width one grid node. The quality of the linearization varies across the lateral domain - it is better with dense sampling in time. It is important to remember that the choice of linearization is irrelevant for the results, it only influences the CPU time of the sampling algorithm.

3.3 RESULTS

The posterior is sampled using the block directional MH algorithm discussed above. We denote 144 updates as one iteration, i.e. on average each grid node is (proposed) updated about once in each iteration. Figure 4 shows trace plots for the log elastic properties at one location in the grid. The traceplots are explained to some extent by the bimodal proposal density q (see Figure 2). In Figure 4 the variables move short distances at some iterations, while moves are large at other iterations, reflecting the bimodal density for the proposal.

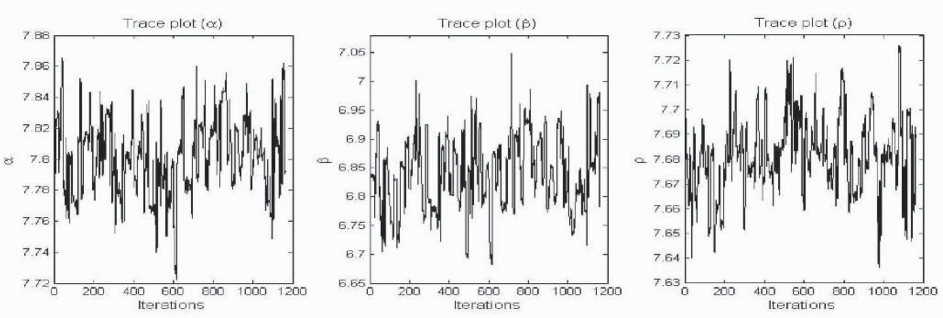


Figure 4. Trace plots of the three variables at one location in the grid. Log of pressure wave velocity α (left), log of shear wave velocity β (middle), and log of density (right).

In Figure 5 we show the estimates of the marginal mean and standard deviation for all three variables as images. Near grid coordinate (North,East) $= (60, 80)$ (see Figure 5, top left image) both pressure and shear wave velocities are large. In the same area the reflection data (Figure 2) are large at both angles. Going south from gridnode $(60, 80)$ (see Figure 5, top left image) the pressure wave velocity decreases, and so does the shear wave velocity, but to a smaller degree. In Figure 2 the reflection data become smaller in this area. These two regions comprise the lobe of the turbidite structure, see Avseth et al (2001). In Eidsvik et al (2004) these two regions were estimated to be water and oil saturated sands, respectively. Without moving on to classifying the velocity and density values, we merely note that pressure wave velocity is larger in water than oil saturated sands (Mavko, Mukerji, and Dvorkin, 1998). Our estimated velocities are hence in accordance with the results in Eidsvik et al (2004). The western part of the domain were predicted to contain mostly shales (a low velocity rock) in Eidsvik et al (2004).

The prior standard deviations for the three variables are $(0.06, 0.11, 0.02)$. In Figure 5 (right) the mean standard deviations in the posterior are $(0.033, 0.065, 0.02)$. This indicates that there is information about α and β in the AVO data (standard deviation decreases by a factor two), but not much about ρ.

Note that the standard deviations in Figure 5 (right) varies quite a lot across the field (a factor of two). The standard deviation for β is smaller where the velocities large. For a linear model [Buland, Kolbjørnsen, and Omre (2003)] the

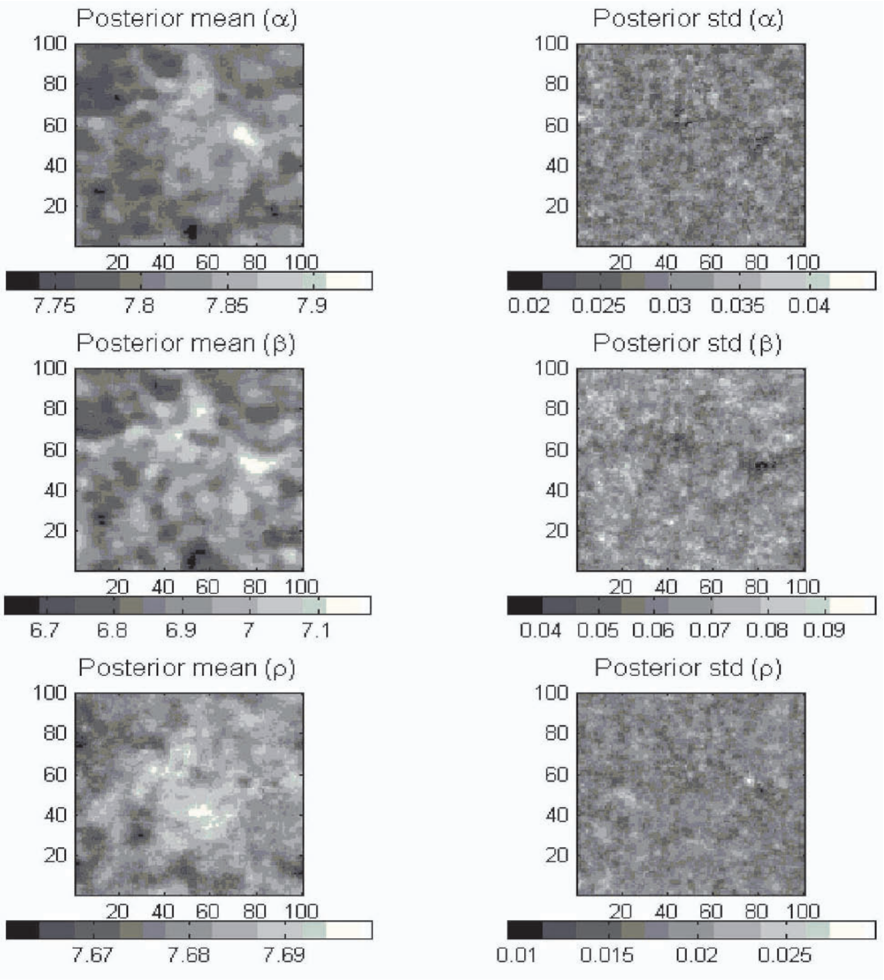


Figure 5. Mean and standard deviation of the three variables at each location in the grid. Logarithm of pressure wave velocity α (top). Logarithm of shear wave velocity β (middle), and logarithm of density (bottom).

standard deviations are constant across the field. The expected values also differ somewhat between a linear model and our nonlinear model; for example $E(\frac{\beta}{\alpha})$ is shifted significant between the two approaches. These differences suggest that the linearized Gaussian posterior in Buland, Kolbjørnsen, and Omre (2003) might introduce a bias in the estimation of the elastic parameters. One might want to correct for the possible bias or variance effects of a linear model, now that this effect is quantified by nonlinear sampling.

4 Closing Remarks

In this paper we consider Bayesian models with a Gaussian random field prior and nonlinear likelihood functions. Such models are common in the earth sciences, but are usually simplified (linearized) to make the posterior analytically available. We propose a directional block Metropolis–Hastings sampler for exploring the original nonlinear posterior. When we apply our methods to a seismic dataset from the North Sea, we recognize some differences between our results and the ones obtained by a linearized model. These differences indicate that it is useful to check the validity of a simplified likelihood model by sampling the full nonlinear models.

One of the current challenges with the Glitne field is uncertainty in the thickness of the turbidite structure, associated with the noise in seismic data due to overburden effects. A natural extension is hence to study the full 3D seismic data. An extension of our statistical methods is to assign priors to the hyperparameters in the statistical model, and hence include the variability of these parameters into the final results.

References

Avseth, P., Mukerji, T., Jørstad, A., Mavko, G., and Veggeland, T., *Seismic reservoir mapping from 3-D AVO in a North Sea turbidite system*, Geophysics, vol. 66, no. 4, 2001, p. 1157-1173

Bassiouni, Z., *Theory, measurement, and interpretation of well logs*, Society of Petroleum Engineers, 1994

Buland, A., Kolbjørnsen, O., and Omre, H., *Rapid spatially coupled AVO inversion in the Fourier domain*, Geophysics, vol. 68, no. 3, 2003, p. 824-835

Cressie, N.A.C., *Statistics for spatial data*, Wiley, 1991

Eidsvik, J., and Tjelmeland, H., *On directional Metropolis–Hastings algorithms*, Submitted for publication, Technical Report, Department of Mathematical Sciences, Norwegian University of Science and technology, http://www.math.ntnu.no/preprint/statistics/2003/S6-2003.pdf

Eidsvik, J., Avseth, P., Omre, H., Mukerji, T., and Mavko, G., *Stochastic reservoir characterization using pre-stack seismic data*, Geophysics, v. 69, no. 4, 2004, p. 978-993

Hegstad, B. K., and Omre, H., *Uncertainty in Production Forecasts based on well observations, seismic data and production history*, Society of Petroleum Engineering Journal, December 2001, p. 409-424

Mavko, G., Mukerji, T., and Dvorkin, J., *The Rock Physics Handbook*, Cambridge, 1998

Pukkila, T.M., and Rao, C.R., *Pattern recognition based on scale invariant functions*, Information Sciences, v. 45, 1988, p. 379-389

Robert, C.P., and Casella, G., *Monte Carlo Statistical Methods*, Springer, 1999

Sheriff, R.E., and Geldart, L.P., *Exploration seismology*, Cambridge, 1995

DETECTION OF LOCAL ANOMALIES IN HIGH RESOLUTION HYPERSPECTRAL IMAGERY USING GEOSTATISTICAL FILTERING AND LOCAL SPATIAL STATISTICS

PIERRE GOOVAERTS
BioMedware, Inc. 516 North State Street, Ann Arbor, MI 48104

Abstract. This paper describes a methodology to detect patches of disturbed soils in high resolution hyperspectral imagery, which involves successively a multivariate statistical analysis (principal component analysis, PCA) of all spectral bands, a geostatistical filtering of regional background in the first principal components using factorial kriging, and finally the computation of a local indicator of spatial autocorrelation to detect local clusters of high or low reflectance values as well as anomalies. The approach is illustrated using one meter resolution data collected in Yellowstone National Park. Ground validation data demonstrate the ability of the filtering procedure to reduce the proportion of false alarms, and its robustness under low signal to noise ratios. By leveraging both spectral and spatial information, the technique requires little or no input from the user, and hence can be readily automated.

1 Introduction

Spatial data are periodically collected and processed to monitor, analyze and interpret developments in our changing environment. Remote sensing is a modern way of data collecting and has seen an enormous growth since launching of modern satellites and development of airborne sensors. In particular, the recent availability of high spatial resolution hyperspectral (HSRH) imagery offers a great potential to significantly enhance environmental mapping and our ability to model spatial systems (Aspinall *et al*., 2002; Marcus, 2002). Following Jacquez *et al.* (2002), HSRH images refer to images with resolutions of less than 5 meters and including data collected over 64 or more bands of electromagnetic radiation for each pixel.

High spatial resolution imagery contains a remarkable quantity of information that could be used to analyze spatial breaks (boundaries), areas of similarity (clusters), and spatial autocorrelation (associations) across the landscape. This paper addresses the specific issue of soil disturbance detection, which could indicate the presence of land mines or recent movements of troop and heavy equipment. A challenge presented by soil detection is to retain the measurement of fine-scale features (i.e. mineral soil changes, organic content changes, vegetation disturbance related changes, aspect changes) while still covering proportionally large spatial areas. An additional difficulty is that no ground data might be available for the calibration of spectral signatures, and little might be known about the size of patches of disturbed soils to be detected. Precise

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 105-114.

and accurate soil disturbance identification typically requires: (1) identification of a potential target (soil disturbance) of interest, (2) removal of confusion (the environmental setting), and (3) target (soil disturbance) confirmation. These different steps should be automated as much as possible to allow for the fast processing of multiple images, while false positives should be reduced to a manageable level.

A major challenge facing the use of HSRH data is the development of new, spatially explicit tools that exploit both the spectral and spatial dimensions of the data. Semivariograms allow one to detect multiple scales of spatial variability, and the spectral values can then be decomposed into the corresponding spatial components using factorial kriging (Goovaerts, 1997; Wackernagel, 1998). This technique has first been used in geochemical exploration to distinguish large isolated values (pointwise anomalies) from groupwise anomalies that consist of two or more neighboring values just above the chemical detection limit (Sandjivy, 1984). Ma and Royer (1988) have applied the same technique to image restoration, filtering and lineament enhancement, while Wen and Sinding-Larsen (1997) have analyzed sonar images. More recently, Van Meirvenne and Goovaerts (2002) applied factorial kriging to the filtering of multiple SAR images, strengthening relationships with land characteristics, such as topography and land use. None of these studies has however tackled the issue of automatic analysis and processing of large series of correlated spectral bands.

This paper describes a new approach that combines geostatistical filtering with local cluster analysis used in health sciences for the detection of clusters and outliers in cancer mortality rates (Jacquez and Greiling, 2003). The methodology is applied to HSRH imagery collected in Yellowstone National Park, and performances are assessed using ground data. Sensitivity analysis is conducted to investigate the impact of spectral resolution, signal to noise ratio, and kernel detection size on classification accuracy.

2 Geostatistical Methodology

Consider the problem of detecting, across an image, single or aggregated pixels that are significantly different from the surrounding ones. The information available consists of *K* variables (i.e. original spectral values or combinations of those) recorded at each of the *N* nodes of the image, $\{z_k(\mathbf{u}_i),\ i=1,\ldots,N;\ k=1,\ldots,K\}$. The proposed approach proceeds in two steps:

1. The regional variability (i.e. spatial background) of the image is filtered in order to highlight local anomalies, which are values that depart from the surrounding mean.
2. At each location across the filtered image the value of a detection kernel, whose size corresponds to the expected size of a patch of disturbed soil, is compared to neighborhood values and flagged as anomaly if its value is significantly higher or lower than surrounding pixel values.

2.1 GEOSTATISTICAL FILTERING

The first step involves removing from each image (i.e. original spectral bands or principal components) the low-frequency component or regional variability. For the *k*-th image, the low-frequency component, denoted m_k, is estimated at each location $\boldsymbol{u}$ as a linear combination of the n surrounding pixel values:

$$m_k(\mathbf{u}) = \sum_{i=1}^{n} \lambda_{ik} z_k(\mathbf{u}_i) \quad \text{with} \quad \sum_{i=1}^{n} \lambda_{ik} = 1 \tag{1}$$

where λ_{ik} is the weight assigned to the i-th observation in the filtering window of size n. The main feature of this filtering technique is that the weights λ_{ik} are tailored to the spatial pattern of correlation displayed by each image and assessed using the semivariogram. These weights are computed as solution of the following system of linear equations (kriging of the local mean):

$$\begin{aligned} \sum_{j=1}^{n} \lambda_{jk}\gamma_k(\mathbf{u}_i\text{-}\mathbf{u}_j) + \mu(\mathbf{u}) = 0 \qquad i = 1,\dots,n \\ \sum_{j=1}^{n} \lambda_{jk} = 1 \end{aligned} \tag{2}$$

where $\gamma_k(\boldsymbol{u}_i\text{-}\boldsymbol{u}_j)$ is the semivariogram of the k-th image for the separation vector between $\boldsymbol{u}_i$ and $\boldsymbol{u}_j$, and $\mu(\boldsymbol{u})$ is a Lagrange multiplier that results from minimizing the estimation variance subject to the unbiasedness constraint on the estimator.

2.2 DETECTION OF ANOMALIES USING THE LISA STATISTIC

The second step amounts at scanning each filtered image, looking for local values that are significantly lower or higher than the surrounding values and might indicate the presence of disturbed soils. This procedure requires the definition of:

1. Detection kernel, whose size corresponds to the expected size of a patch of disturbed soil,
2. LISA neighborhood including the pixels surrounding the detection kernel,
3. Target area which is the area to be analyzed.

An example of these three parameters is provided in Figure 1.

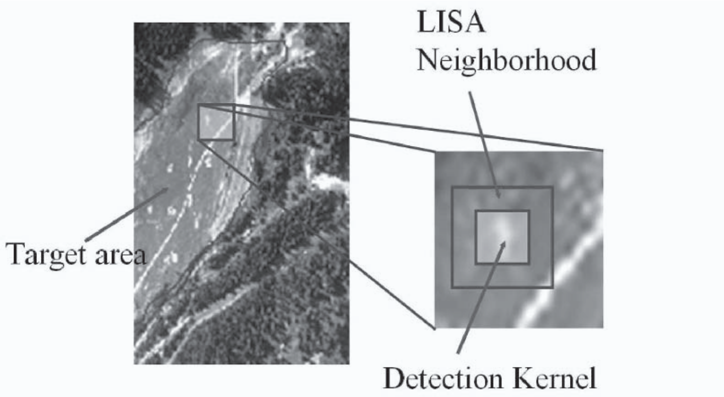


Figure 1. Illustration of key parameters used in the detection procedure.

The detection of local anomalies is based on the local Moran's I, which is the most commonly used LISA (Local Indicator of Spatial Autocorrelation) statistic (Anselin, 1995). It is computed for each pixel of coordinates ***u*** as:

$$\text{LISA}_k(\mathbf{u}) = \bar{r}_k(\mathbf{u})\left[\frac{1}{J}\sum_{i=1}^{J} r_k(\mathbf{u}_i)\right] \tag{3}$$

where $\bar{r}_k(\mathbf{u})$ is the average value of the residuals, $r_k(\boldsymbol{u})=z_k(\boldsymbol{u})-m_k(\boldsymbol{u})$, over the detection kernel centered on pixel of coordinates ***u***, and J is the number of pixels in the LISA neighborhood (e.g. J=12 and kernel comprises 4 pixels for the example of Figure 1). Since the residuals have zero mean, the LISA statistic takes negative values if the kernel average is much lower (or higher) than the surrounding values. In other words the kernel average is below the global zero mean while the neighborhood average is above the global zero mean, or conversely, which indicates the presence of anomalies. Clusters of low or high values will lead to positive values of the LISA statistic (e.g. both kernel and neighborhood averages are jointly above zero or below zero).

In addition to the sign of the LISA statistic, its magnitude informs on the extent to which kernel and neighborhood values differ. To test whether this difference is significant or not, a Monte Carlo simulation is conducted, which consists in sampling randomly the target area and computing the corresponding simulated neighborhood averages. This operation is repeated many times (e.g. 1,000 draws) and these simulated values are multiplied by the detection kernel average $\bar{r}_k(\mathbf{u})$ to produce a set of 1,000 simulated values of the LISA statistic at ***u***. This set represents a numerical approximation of the probability distribution of the LISA statistic at ***u*,** under the assumption of spatial independence. The observed LISA statistic, LISA$_k(\boldsymbol{u})$, can then be compared to the probability distribution, allowing the computation of the probability that this observed value could be exceeded (so-called p-value):

$$p_k(\mathbf{u}) = \text{Prob}\{L > LISA_k(\mathbf{u}) \mid \text{randomization}\} \tag{4}$$

Large p-values thus indicate large negative LISA statistic, corresponding to small values surrounded by high values or the reverse (anomalies or presence of negative local autocorrelation). Conversely, small p-values correspond to large positive LISA statistic which indicates clusters of high or low values (positive autocorrelation).

The last step is to combine the K p-values computed for the set of K images. Two novel statistics were developed to summarize for each node ***u*** the information provided by the K bands and to detect target pixels:

1. Average p-value over the subset of K' bands that display negative LISA statistic:

$$S_1(\mathbf{u}) = \frac{1}{K'}\sum_{k=1}^{K} i(\mathbf{u};k)\,p_k(\mathbf{u}) \qquad \text{and} \qquad K' = \sum_{k=1}^{K} i(\mathbf{u};k) \tag{5}$$

with $i(\boldsymbol{u};k)$ = 1 if LISA$_k(\boldsymbol{u})$ < 0, and zero otherwise. Large S_1 values indicate local anomalies (i.e. sample LISA statistic in the left tail of the distribution).

2. Average absolute deviation of p-values from 0.5 through the K bands:

$$S_2(\mathbf{u}) = \frac{1}{K}\sum_{k=1}^{K} | p_k(\mathbf{u}) - 0.5 | \tag{6}$$

   Large S_2 values indicate either clusters or anomalies (i.e. sample LISA in either tails of the distribution).

The detection procedure requires applying a threshold to the maps of statistics S_1 or S_2 and classifying as disturbed soils all pixels exceeding this probability threshold. Instead of selecting a single threshold arbitrarily, it is better to select a series of thresholds and see how the proportion of pixels correctly or incorrectly classified as disturbed soils evolves. This information can then be summarized in the so-called Receiver Operating Characteristics (ROC) curve that plots the probability of false alarm versus the probability of detection.

3 Case study

The new methodology was tested on a vegetation plot located in the northern boundary area of Yellowstone National Park. The objective is to detect 4 blue tarps of 4m^2 area in the image (131×69 pixels). These four targets mainly correspond to the white pixels in the image of Figure 2 (left), and are denoted by the black squares in the right image. These data were collected using the Probe-1 sensor, a 128-band hyperspectral system operated by Earth Search Systems, Inc. To obtain 1 m resolution data, this sensor was mounted on a helicopter flying approximately 600 m above the ground. Following atmospheric correction, the images were degraded in order to investigate the robustness of the approach with respect to spatial resolution and signal to noise ratio. The data were first spectrally resampled to 2-5 times lower resolutions, by simply selecting one out of every 2 to 5 bands. Noise was added to simulate 50:1 signal-to-noise ratio (SNR) and 100:1 SNR, according to: $R_{sn}(\lambda)=R_s(\lambda)[1+\{N(0,1)/SNR(\lambda)\}]$, where $R_{sn}(\lambda)$ is the simulated, noisy spectrum, $R_s(\lambda)$ is the spectrum that has been spectrally resampled, N(0,1) is a Gaussian random number with a zero mean and unit variance, and SNR(λ) is the simulated signal-to-noise ratio.

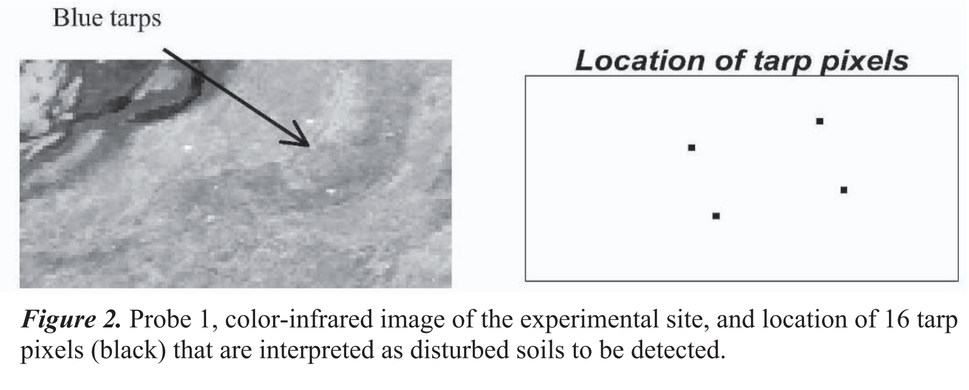


Figure 2. Probe 1, color-infrared image of the experimental site, and location of 16 tarp pixels (black) that are interpreted as disturbed soils to be detected.

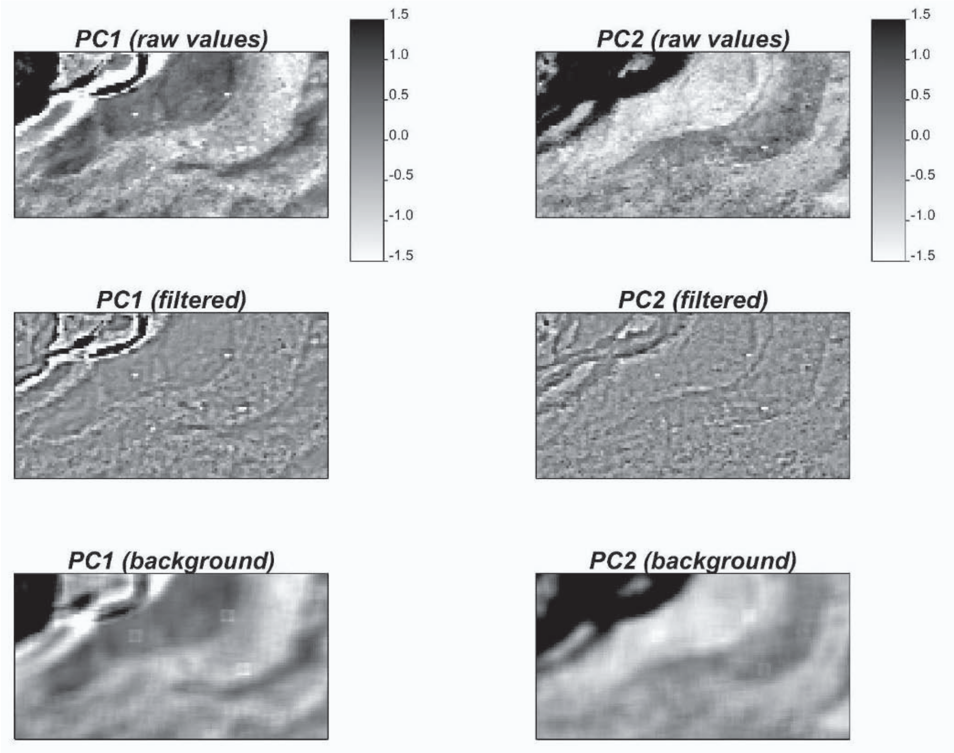


Figure 3. Maps of the first two principal components for the HSRH image, and the results of the geostatistical filtering of the regional background.

The analysis was first performed on the first 84 principal components (PC) of the data. Each image of principal components was decomposed into maps of local means and residuals or filtered values, using a 5×5 window centered on the pixel being filtered (i.e. n=25 in equation (1)). Figure 3 shows an example for the first 2 PCs. The original PC values are decomposed into the background values $m(\mathbf{u})$ and the residuals or filtered values $r(\mathbf{u})=z(\mathbf{u})-m(\mathbf{u})$. These images illustrate how the removal of regional variability, which might represent different soil or vegetation types, highlights the location of target pixels which appear as white in the filtered images. The information provided by either filtered or non filtered sets of 84 PCs was summarized using the statistic S_1 or S_2, see Figure 4. Dark pixels, corresponding to high values, indicate the presence of local anomalies for S_1 and clusters or anomalies for S_2. This figure clearly illustrates the benefit of the geostatistical filtering and use of statistic S_2, which reduces greatly the number of background pixels being wrongly detected as clusters or local anomalies and increases the similarity with the actual map of tarp pixels displayed in Figure 2.

The final step is to compute the ROC curves from the maps of statistic S_1 or S_2. A series of thresholds (probability of detection) are selected, and for each of them the pixels classified as disturbed soils are compared to ground data in order to compute the proportion of misclassified pixels (probability of false alarms). These two sets of probabilities are then plotted to generate the ROC curve. Figure 5 shows an example of such curves for detection using either statistic S_1 or S_2 computed from filtered or non-filtered images. The main conclusions are:

1. The filtering and use of statistic S_2 (black solid curve) allows the detection of all tarp pixels for a probability of false alarms not exceeding 0.20.
2. Detection of 60% of tarp pixels can be done with small probability of false alarm (vertical part of the ROC curve) and these pixels correspond to high purity in term of tarp content. Pixels that contain a mixture of tarp and other materials (i.e. bare soil, grass) are much more difficult to detect and generate an increase in the proportion of false alarms which can be fairly dramatic if no filtering is performed and only anomalies are searched (i.e. use of statistic S_1).

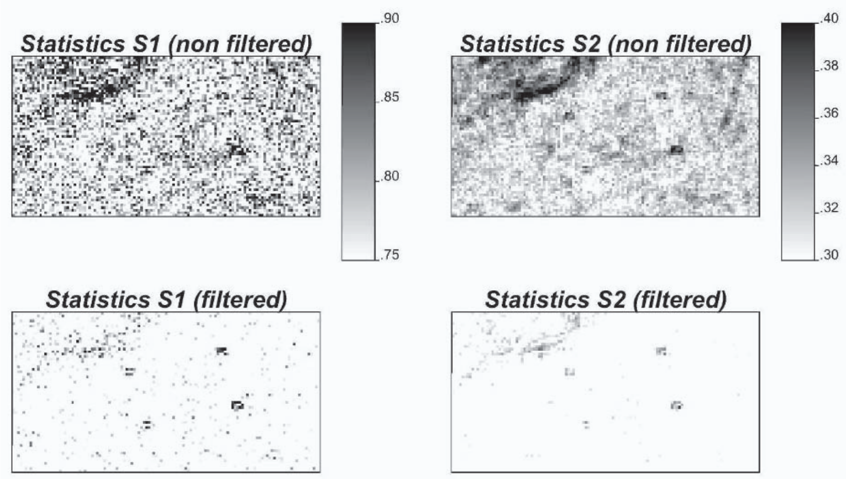


Figure 4. Maps of statistics S_1 and S_2 computed from the first 84 principal components before and after (bottom maps) filtering of the regional background.

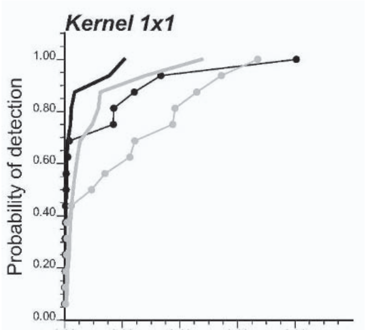


Figure 5. Receiver Operating Characteristics (ROC) curves obtained for the statistics S_1 (thin dotted line) and S_2 (solid line). Black curves are obtained from the filtered values, while the gray curves refer to original values (without geostatistical filtering).

Extensive sensitivity analysis has been conducted to assess the performance of the methodology under several conditions, such as:

1. Selection of subsets of PCs based on the strength of spatial correlation.
2. Choice of detection kernels of various sizes.
3. Decrease in signal to noise ratio and spectral resolution.

Instead of summarizing the information provided by the first 84 PCs, statistics S_1 and S_2 were computed for each PC separately and their average for both tarp and background pixels are plotted versus the rank/order of the PC in Figure 6 (top graph). Differences between tarp (black) and background (gray) pixels tend to attenuate as the order of the component increases and the spatial correlation of the image decreases (thick black curve). Subsets of PCs were thus retained based on a spatial correlation threshold of 0.5 or 0.25, plus the set of the first 25 PCs. The ROC curves indicate an increase in the proportion of false alarms when using fewer PCs. All ROC curves computed hereafter will be based on the first 25 PCs, thereby providing a balance between shorter CPU time (16.0 seconds versus 54.5 for 84 PCs on a Pentium 3.20 GHz) and slightly more false alarms.

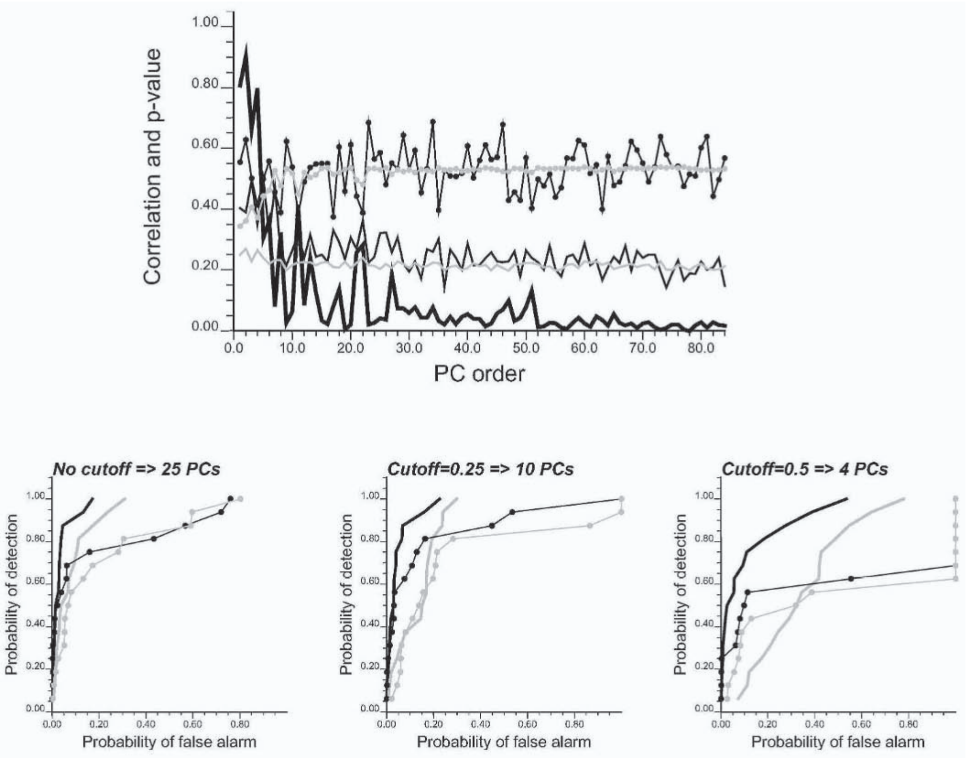


Figure 6. Plot of spatial correlation (lag=1 pixel) and value of statistics S_1 (thin dotted line) and S_2 (solid line) for either tarp pixels (black) or background pixels (gray), versus the order of the principal component (top graph). Bottom graphs show the ROC curves obtained for the first 25 PCs and two subsets based on the level of spatial correlation.

All results presented so far were obtained using a detection kernel of one pixel, which does not require any prior information regarding the size of the object to be detected. The benefit of tailoring the detection kernel to the size of the object was investigated by performing the classification and computing the ROC curves for three types of kernel: 1×1, 2×1 and 2×2. Figure 7 (top row) shows that that the use of kernels 2×1 and 2×2 improves detection performances of statistic S_1, while more false alarms occur when using statistic S_2. Indeed, statistic S_1 searches for local anomalies of size equal to the kernel, while S_2 detects both clusters and anomalies. The impact of the signal to noise (SN) ratio was investigated by adding a given proportion of noise to reflectance values before performing PCA. Figure 7 (middle row) shows the ROC curves obtained for increasing levels of noise (SN=100:1 to SN=50:1). As intuitively expected, noisy signals tend to blur the detection of anomalies, leading to a larger proportion of false alarms, although statistic S_2 on filtered signal is very robust.

The last test consisted in investigating how a decrease in spectral resolution would affect the quality of the detection. Figure 7 (bottom row) shows the ROC curves obtained for the original signal with 84 PCs, and then for one half (WV2, 42 PCs) and one third (WV3, 28 PCs) of the number of principal components. As for the signal to noise ratio, ROC curves indicate poorer performances when using the degraded image.

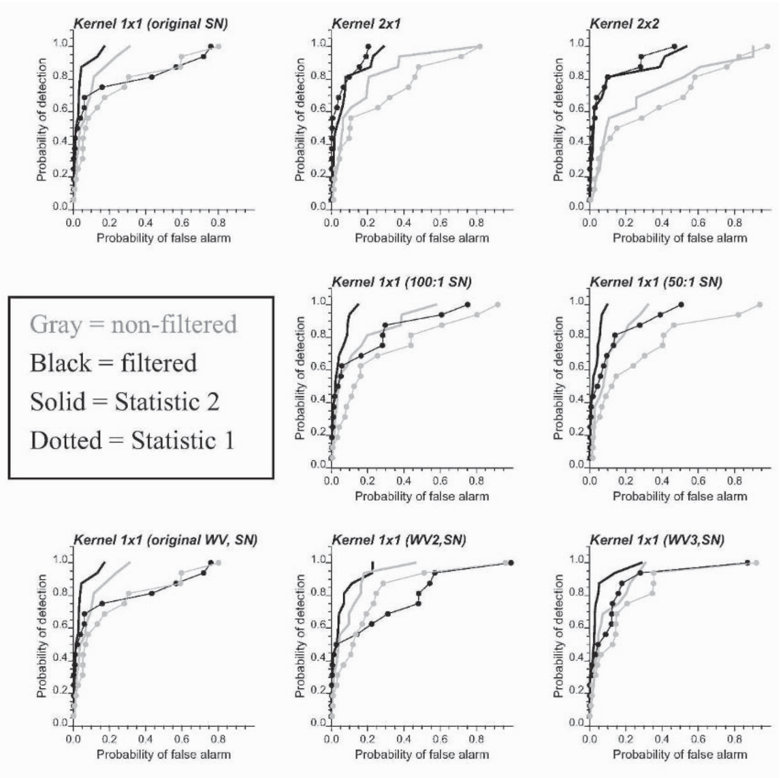


Figure 7. Receiver Operating Characteristics (ROC) curves obtained for three types of detection kernel, two signal to noise (SN) ratios, and three spectral resolutions (WV).

4 Conclusions

This paper presented and demonstrated the efficacy of a geostatistical approach to detecting disturbed soils in high spatial resolution hyperspectral imagery. The technique uses PCA to reduce dimensionality of the imagery, employs geostatistical filtering to remove regional background and enhance local signal, and applies a Local Indicator of Spatial Autocorrelation to identify patches of disturbed soils. In all scenarios, fewer false alarms were obtained when using the filtered signal and statistic S_2 to summarize information across bands. Image degradation through addition of noise or reduction of spectral resolution tends to blur the detection of anomalies, leading to more false alarms, in particular for the identification of the few mixed pixels.

In this paper the methodology was used to detect regular patches on a simple landscape. Similar results were obtained when applying the approach to more complex landscapes with multiple targets of various sizes and shapes (results not shown). Because it employs geostatistical filtering, the method is robust under low signal to noise ratios. By leveraging both spectral and spatial information, the technique requires little or no input from the user, and hence can be readily automated. Following our results a Pentium 3.20 GHz would allow the processing of a 1000×1000 scene including 25 bands within 18 minutes. Future research will investigate the benefit of processing directly the spectral bands instead of their principal components.

Acknowledgements

This work was supported by NAVAIR SBIR Phase I N02-172, and data were kindly provided by Andrew Marcus from the University of Oregon.

References

Anselin, L., Local indicators of spatial association-LISA, *Geographical Analysis*, vol. 27, 1995, p. 93-115.

Aspinall, R.J., Marcus, W.A., and Boardman, J.W., Considerations in collecting, processing, and analysing high spatial resolution hyperspectral data for environmental investigations, *Journal of Geographical Systems*, vol. 4, 2002, p. 15-29.

Goovaerts, P., *Geostatistics for Natural Resources Evaluation*, Oxford University Press, 1997.

Jacquez, G.M. and Greiling, D.A., Local clustering in breast, lung and colorectal cancer in Long Island, New York, *International Journal of Health Geography*, vol. 2, no. 3, 2003.

Jacquez, G.M., Marcus, W.A., Aspinall, R.J. and Greiling, D.A., Exposure assessment using high spatial resolution hyperspectral (HSRH) imagery, *Journal of Geographical Systems*, vol. 4, 2002, p. 15-29.

Ma, Y.Z., and Royer, J.J., Local geostatistical filtering : application to remote sensing, *Sciences de la Terre, Serie Informatique,* vol. 27, 1988, p. 17-36.

Marcus, W.A., Mapping of stream microhabitats with high spatial resolution hyperspectral imagery, *Journal of Geographical Systems*, vol. 4, 2002, p. 113-126.

Sandjivy, L., The factorial kriging analysis of regionalized data. Its application to geochemical prospecting. In *Geostatistics for Natural Resources Characterization*, edited by G. Verly, M. David, A.G. Journel, and A. Marechal, Dordrecht: Reidel, 1984, p. 559-571.

Van Meirvenne, M. and Goovaerts, P., Accounting for spatial dependence in the processing of multitemporal SAR images using factorial kriging, *International Journal of Remote Sensing*, vol. 23, 2002, p. 371-387.

Wackernagel, H., *Multivariate Geostatistics*, Springer-Verlag, 1998.

Wen, R. and Sinding-Larsen, R., Image filtering by factorial kriging—sensitivity analysis and application to Gloria side-scan sonar images, *Mathematical Geology*, vol. 29, no. 5, 1997, p. 433-468.

COVARIANCE MODELS WITH SPECTRAL ADDITIVE COMPONENTS

DENIS MARCOTTE and MIRO POWOJOWSKI
Département des génies civil, géologique et des mines,
École Polytechnique, C.P. 6079, Succ. Centre-ville, Montréal, QC, H3C 3A7

Abstract. We present a new model defining a whole class of variogram models: the spectral additive model (SAM). The model is obtained by linear combination of simple spectral components. The SAM parameters can be estimated linearly and without bias. The handling of mean drift is straightforward. In the spatial domain, the SAM possesses an analytic expression, a clear advantage over similar approaches based on covariance spectra obtained by FFT. The SAM is flexible as it can approximate any classical model, isotropic or anisotropic, to the desired degree of precision. A forward inclusion selection procedure enables avoiding over-parameterization of the model. This is especially useful in the general anisotropic case. Simulations illustrate the performance of the SAM for covariance function fitting.

1 Introduction

The choice of a suitable variogram or covariance model is an important step in any geostatistical study. This step remains largely handcrafted and resists automation. Current practice normally involves computing experimental variogram(s); a step involving its legion of more or less arbitrary decisions like the choice of directions, the angular tolerance, and the distance bins to adopt. Decisions on the characteristics of the model follow: stationary or non-stationary, isotropic or anisotropic, type of anisotropy, type of model or of combination of models to use. Finally the model parameters are adjusted, either manually or sometimes with the help of automatic fitting programs and cross-validation procedures (Marcotte, 1995). Most of the classical models being non-linear functions of distance, automatic fitting itself can be difficult to realize as many local optimums could exist. This partly explains why the parameters are still often obtained by visual fit.

Although a host of models are available (Chilès and Delfiner, 1999), one may question why data should necessarily comply with any of these models. There is a definite need to introduce greater flexibility and ease of estimation, especially if one considers implementation of geostatistical algorithms in wide general use software packages like GIS and statistical packages. The need for automation and flexibility is certainly present in the univariate stationary or non-stationary cases but it is even more compelling in the muzltivariate case.

We present a new class of models that are flexible and suitable for automatic estimation. This class of models is obtained by linear combinations of spectral components, thus the name spectral additive model (SAM). Because the model is linear, its parameters can be estimated by standard regression (or robust or weighted

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 115-124.

regression if preferred). There is no need to compute a variogram for the estimation because the fitting can be done easily considering each data pair available. The focus of this study is on the univariate case, stationary or non-stationary, isotropic or anisotropic.

The proposed approach bears resemblance to the approach suggested by Yao and Journel (1998). However, it is fundamentally different in many aspects. Here, the model defined is continuous and an analytical expression for the covariance function exists. In Tiao and Journel (1998) the model is discrete and numerically defined only at the lags used in the variogram computation.

2 Theory

2.1 A CLASS OF FLEXIBLE ISOTROPIC MODELS

Stationary (or homogeneous) random functions with absolutely integrable covariance function *C(h)* possesses the spectral density $c(\omega)$. Together, they form a Fourier transform pair (Christakos, 1992, Yaglom, 1987). That is,

$$\begin{aligned} &C(h) \xrightarrow{\Im} c(\omega) \\ &C(h) \xleftarrow[\Im^{-1}]{} c(\omega) \end{aligned} \tag{1}$$

with $c(\omega)>0$ for all frequencies ω and $c(\omega)$ symmetric.

The basic idea in our approach is to replace the continuous function $c(\omega)$ by a summation of piecewise continuous functions $c_i(\omega)$, that we call spectral components:

$$c_i(\omega) = f_i(|\omega|)\ 1_{r_{i-1}\leq|\omega|<r_i} \qquad \omega \in \Re^d \tag{2}$$

with $0=r_0<r_1<...r_n\leq +\infty$ being an increasing finite sequence of positive numbers, and $1_{r_{i-1}\leq|\omega|<r_i}$ being an indicator function. The bins $r_0, r_1,...r_n$ are selected so as to provide good coverage of the positive frequencies. Although there is freedom in the choice of $f_i(\omega)$, a simple and convenient choice is $f_i(\omega)=1$ for all intervals except possibly for the last semi-infinite interval.

Now consider a linear combination of the spectral components:

$$c(\omega) = \sum_{i=1}^{n} a_i c_i(\omega) \tag{3}$$

Its Fourier transform is:

$$C(h) = \sum_{i=1}^{n} a_i \Im^{-1}(c_i(\omega)) = \sum_{i=1}^{n} a_i C_i(h) \tag{4}$$

The following results hold true:

R1. Any linear combination with coefficients $a_i \geq 0$ is the spectral density of an admissible covariance;
R2. Any admissible isotropic covariance having a spectral density can be approximated to an arbitrary degree of accuracy by model (4) (Powojowski, 2000). The accuracy of the approximation increases with the number of spectral components used to discretize the spectral density.

The isotropic covariance corresponding to the choice $f_i(\omega) = 1$ in Equation 2 is given by:

$$C_i(h) = h^{-d/2}(r_i^{d/2} J_{d/2}(r_i h) - r_{i-1}^{d/2} J_{d/2}(r_{i-1} h)) \qquad (5)$$

where $J_{d/2}$ is the order d/2 Bessel function of the first kind, where d is the dimension of the space. Note that for the limit where h->0:

$$C_i(0^+) = (r_i^{d/2} - r_{i-1}^{d/2})/\Gamma(1 + d/2) \qquad (6)$$

Figure 1 shows the isotropic covariance for a few selected spectral components after normalisation by $C_i(0^+)$ (Equation (6)) to ensure a unit sill.

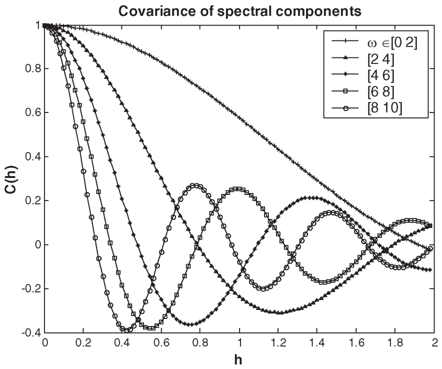


Figure 1. Equation 5 and normalized to a unit sill.

2.2 CONTROLLING MODEL BEHAVIOUR AT THE ORIGIN

In some cases, it is convenient to choose the last interval semi-open to infinity. This interval controls the behaviour of the covariance function at the origin. For example, the following choice ensures linear behavior at the origin:

$$c_n(\omega) = g(|\omega|)\, 1_{r_n < |\omega|} \qquad \omega \in \Re^d \qquad (7)$$

where $g(\omega)$ is the spectral density of the isotropic exponential covariance. When d=2, $g(|\omega|) = \frac{1}{a}\left(|\omega|^2 + \frac{1}{a^2}\right)^{-3/2}$, with a the range. Setting $f_n(\omega) = 0$ gives a differentiable covariance model.

2.3 APPROXIMATION OF CLASSICAL PARAMETRIC COVARIANCE MODELS BY SPECTRAL COMPONENTS : THE ISOTROPIC CASE

Figure 2 shows some 2D-isotropic parametric covariance models and their close approximation by a model with 4 (a and d) or 5 (b and c) spectral components. All spectral components are piecewise constants. However, in cases b) and c) an exponential spectral component is added to reproduce the model linear behavior. Clearly, the spectral additive model can match any isotropic classical model with few spectral components, thus demonstrating its great flexibility.

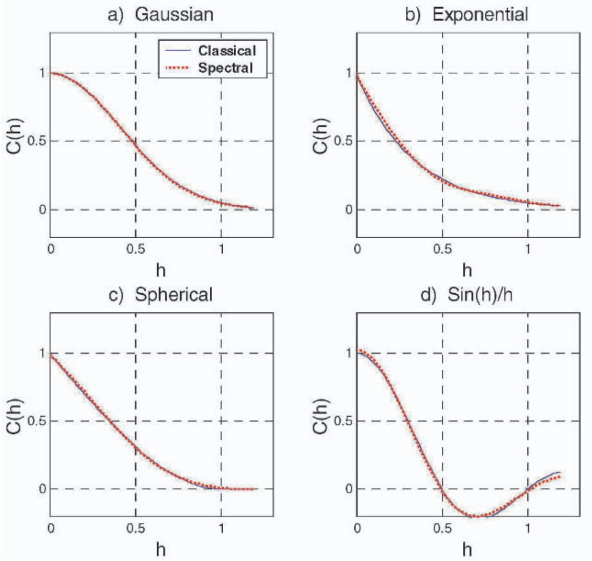


Figure 2. Examples of close approximation of various isotropic covariances by additive spectral models.

2.4 ESTIMATION OF PARAMETERS

A definite advantage of the spectral additive model is the possibility of estimating its parameters linearly. We assume, for generality, a model with an unknown mean that can include a trend.

The drift model is:

$$Z(x) = X\beta + \eta(x) \qquad (8)$$

where $\eta(x)$ is second order stationary with zero mean and real covariance K_Z and X is the nxp regression matrix used to model the drift.

The residuals Y_i, i=1...n, are obtained by ordinary least-squares estimation of β using the n available data:

$$Y = \left(I - X(X'X)^{-1}X'\right)Z = PZ \qquad (9)$$

where Y and Z are vectors of size nx1, I is the identitiy matrix of order n and P is the nxn projection matrix.

The product of residuals is YY'. Its expectation is:

$$E[YY'] = PK_Z P \tag{10}$$

The covariances for the *n* data points computed from SAM are:

$$K_a = \sum_{j=1}^{q} a_j K_j \tag{11}$$

Each matrix K_j is *n*x*n.* It represents the covariances associated with the "jth" spectral component. There are *q* such spectral components.

The covariances for the residuals obtained with SAM are:

$$PK_a P = \sum_{j=1}^{q} a_j PK_j P = \sum_{j=1}^{q} a_j U_j \tag{12}$$

Estimators of a_j minimize the norm between the product of residuals YY' and the covariances computed with Equation 12:

$$min \left\| YY' - \sum_{j=1}^{q} \hat{a}_j U_j \right\|_2 \tag{13}$$

The least squares estimator is:

$$\hat{a} = \left[trace(U_i U_j) \right]^{-1} \left[Y' U_i Y \right] \text{ with } \hat{a}_i \geq 0, i = 1...q \tag{14}$$

where $\hat{a}$ is the *qx1* vector of coefficients for the *q* spectral components. The notation $\left[trace(U_i U_j) \right]$ denotes a *qxq* matrix whose (i,j)- th entry is $trace(U_i U_j)$.

Note that a nugget effect can be included as an additional component in Equations 11 to 14.

2.5 CONSTRAINTS ON THE a_i COEFFICIENTS

For the SAM to be admissible, the coefficients a_i must be non-negative. Enforcing these constraints directly in the estimation procedure complicates the computation. One way to circumvent this problem is to forward select the spectral components one at a time. At each step, the spectral component providing the best-fit improvement to the product of residuals and, at the same time, providing a set of positive coefficients, is selected. The procedure is stopped when no further significant improvement is possible or when all the candidate spectral components provide inadmissible models (i.e. at least one coefficient becomes negative).

2.6 A DIFFERENT NORM

The norm used in Equation 13 gives the same weight to all products of residuals. It is well known that the experimental variogram or covariance function is more reliable at short distances than at large distances due to smaller fluctuations. Also, the covariance at short distances has more influence on geostatistical operators like kriging or simulations. Thus, it could be interesting to favour the covariance fit at short distances by considering instead the modified norm:

$$min\left\| YY'*V-\sum_{j=1}^{q}\hat{a}_jU_j*V\right\|_2 \qquad (15)$$

where $U*V$ is the Hadamard matrix product (i.e. element by element multiplication) and V is a *nxn* weighting matrix with elements defined by a positive non-increasing function of the distance separating the data.

With this weighting, the parameter estimates are now given by:

$$\hat{a}=\left[trace\left((U_i*V)U_j\right)\right]^{-1}\left[Y'(U_i*V)Y\right] \text{ with } \hat{a}_i\geq 0,\ i=1...q \qquad (16)$$

2.7 SELECTION OF SPECTRAL BINS

Shannon’s (1949) sampling theorem for data on a regular grid indicates the highest frequency component that can be estimated reliably from the data. This frequency, the Nyquist frequency, is given by:

$$f_{Nyq}=\frac{1}{2\Delta h} \qquad (17)$$

where Δh is the grid step.

For irregularly spaced data, the Nyquist frequency is not defined. We compute a pseudo-Nyquist frequency using Equation 17 by substituting the grid spacing Δh by the average distance for the 30 nearest data pairs.

At frequencies higher than the Nyquist frequency, an exponential spectral component can be added to the set of piecewise constant spectral components to impose linear behaviour of the covariance function at the origin. Shannon’s sampling theorem indicates this decision is essentially model-based and cannot be derived from the data. This may sound paradoxical as the first points of the experimental variogram show less fluctuation and are often well estimated. Nevertheless, the (non) differentiability of the process, that is the linear or parabolic behavior at the origin remains a modelling decision. In practical terms, the fact that the first variogram points define a straight line does not guarantee it extrapolates linearly to the intercept.

Having defined the highest frequency available, equal bins are used to define the various spectral components. A more elaborate spectral binning strategy is described in Powojowski (2000).

2.8 PROPERTIES OF THE ESTIMATOR

Powojowski (2000) studied in detail the properties of the estimator. He showed that when the true covariance function has the form of Equation 11, the estimator $\hat{a}$ is unbiased. Otherwise $\hat{a}$ is still a meaningful estimator in the sense that it is the closest to *YY*' for the chosen norm. It was shown that, with a known mean, the estimator is convergent under in-fill sampling - expanding domain conditions. However, for a compact domain the estimator has a residual variance even in the case of in-fill sampling. With an estimated mean, convergence is not ensured except if the weights in *V* (Equation 16) decay exponentially with distance.

2.9 A SPECTRAL ANISOTROPIC MODEL

The general methodology described above for isotropic models extends readily to general anisotropic models. The idea is to bin the 2D or 3D frequency domain. To illustrate, we consider only the 2D case. The covariance is an even function: $c(\omega_1,\omega_2) = c(-\omega_1,-\omega_2)$. Thus only half the frequency plane need be considered.

In 2D, to the spectral component:

$$c_{ij}(\omega_x,\omega_y) = 1 \quad \begin{Bmatrix} \omega_{\mathrm{x,i\text{-}1}} < \omega_x < \omega_{x,i} & & -\omega_{\mathrm{x,i\text{-}1}} > -\omega_x > -\omega_{x,i} \\ & \textit{and} & \\ \omega_{\mathrm{y,j\text{-}1}} < \omega_y < \omega_{y,j} & & -\omega_{\mathrm{y,j\text{-}1}} > -\omega_y > -\omega_{y,j} \end{Bmatrix} \qquad (18)$$
$$0 \quad \text{elsewhere}$$

corresponds the anisotropic covariance:

$$C_{ij}(h_x,h_y) = \frac{cos(\omega_{x,i}h_x + \omega_{y,j\text{-}1}h_y) + cos(\omega_{x,i\text{-}1}h_x + \omega_{y,j}h_y)}{(\pi\, h_x h_y)}\ldots$$
$$\frac{-\,cos(\omega_{x,i}h_x + \omega_{y,j}h_y)\,\text{-}\,cos(\omega_{x,i\text{-}1}h_x + \omega_{y,j\text{-}1}h_y)}{(\pi\, h_x h_y)} \qquad (19)$$

The following limits for Equation 19 exist:

$$h_x \to 0;\, C(0^+,h_y) = \frac{(\omega_{x,j} - \omega_{x,j-1})(sin(\omega_{y,i}h_y) - sin(\omega_{y,i-1}h_y))}{\pi\, h_y}$$
$$h_y \to 0;\, C(h_x,0^+) = \frac{(\omega_{y,j} - \omega_{y,j-1})(sin(\omega_{x,i}h_x) - sin(\omega_{x,i-1}h_x))}{\pi\, h_x} \qquad (20)$$
$$h_x,h_y \to 0;\, C(0^+,0^+) = \frac{(\omega_{x,i} - \omega_{x,i-1})(\omega_{y,j} - \omega_{y,j-1})}{\pi}$$

Figure 3 shows the good fit obtained for a few classical models presenting geometric anisotropy. Note that the SAM can accommodate also any kind of zonal anisotropy.

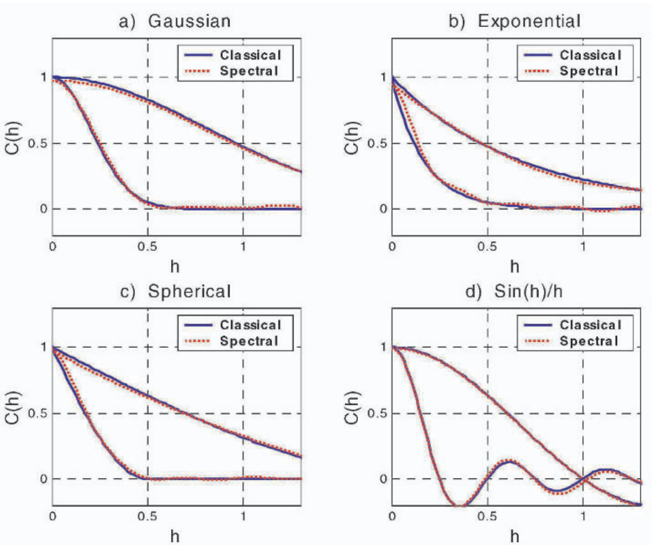


Figure 3. Examples of close approximation of various anisotropic covariances by additive spectral models.

3 Simulated examples

Two hundred data are simulated in a 100m x 100m square. The simulated model is spherical isotropic with a range of 30m, no nugget and a sill of 10. A global linear drift is added with m(x,y)=0.2*x+0.1*y where x and y are the spatial coordinates. Figure 4 shows the experimental covariance, the model covariance and the expected covariance of residuals obtained by fitting an isotropic SAM with drifts of order 0, 1 and 2. Note how the SAM retrieves very well the main characteristics of the simulation for drift orders 1 and 2. On the other hand, when adopting a zero order drift, the model is forced to include strong small frequency components (large range) to account for the drift not included in the model. Although the expected covariances match well the experimental covariances, the theoretical covariances are well above the experimental ones, a clear indication that the variance of the process can not be defined and therefore that the process is non-stationary.

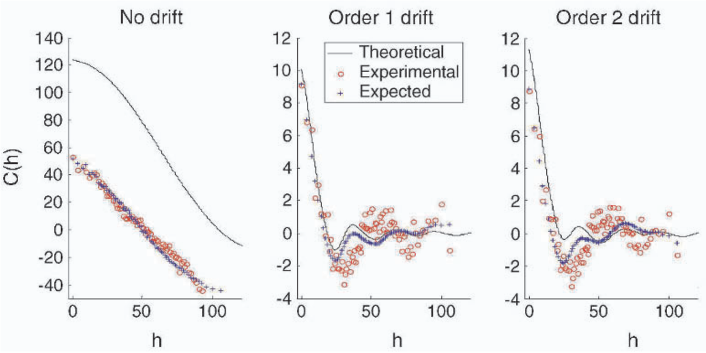


Figure 4. Experimental covariances, expected values of product of residuals and theoretical covariance for the simulated example. Isotropic case.

A similar example is simulated with geometric anisotropy (a_x=50, a_y=25). Figure 5 shows the results obtained with an anisotropic spectral model when specifying order 0 and order 1 drifts. When the right drift order is selected (i.e. order 1), the model correctly identifies the anisotropy present.

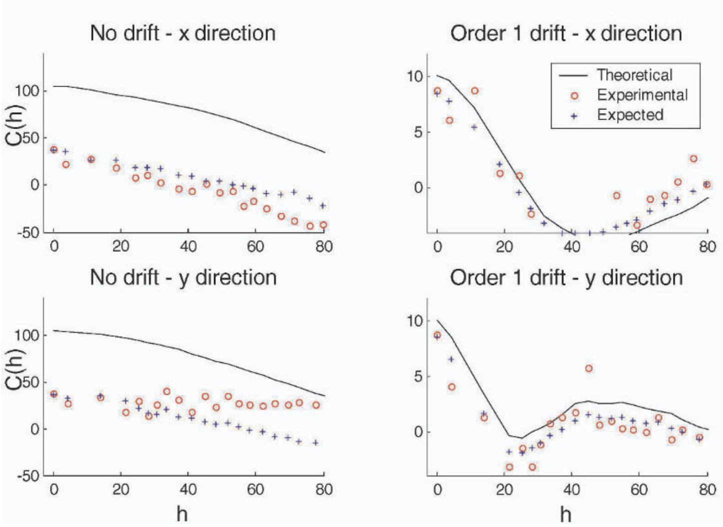


Figure 5. Experimental covariances, expected values of product of residuals and theoretical covariance for the simulated example. Anisotropic case.

5 Discussion and conclusion

The SAM approach reduces covariance model identification to a problem of linear regression with a simple positivity constraint on the regression coefficients. The control variable is the product of residuals. The regressors are the expectation of this product computed from each spectral component (Equation 12). This approach has interesting advantages. First, the class of models so defined is larger than for the usual parametric models. Second, as the parameters can be estimated linearly, it lends itself to automation. Third, all the tools of standard regression can be exploited: statistical tests, identification of outliers, ridge or robust procedures, etc. Fourth, possibly most importantly, it enables estimating the parameters of the model in the presence of mean drift as the effect of the drift is accounted explicitly when computing expected values of product of residuals. Finally, component selection procedures like forward inclusion or backward elimination, or a combination of both, can be used to limit the number of parameters and avoid possible difficulties due to colinearities between the regressors when the number of spectral components is high. In the examples presented, an underestimation of the drift order was easily detected by comparing the theoretical model to the expected value for the product of residuals.

Generalization of the approach to the multivariate case is possible. With “p” variables, “p” real coefficients (one for each spectral density) and p(p-1)/2 complex coefficients (one for each cross-spectral density) will have to be estimated for each spectral bin. The resulting complex coefficient matrix of size pxp must be Hermitian positive semi-definite for the model to be admissible (Wackernagel, 1995).

Acknowledgements

This research was financed by a NSERC grant. We thank Maria Annechione for editing the manuscript. Helpful suggestions from two anonymous reviewers are also acknowledged.

References

Chilès, J. P., and Delfiner, P., 1999, Geostatistics: Modeling spatial uncertainty: Wiley, New York, 695 p.

Christakos, G. 1992. Random Field Models in Earth Sciences. Academic Press, San Diego.

Marcotte, D., 1995. Generalized cross-validation for covariance model selection and parameter estimation. Mathematical Geology, v. 27, no. 6, 749-762.

Powojowski, M., 2000. Sur la modélisation et l'estimation de la fonction de covariances d'un processus aléatoire. Ph. D. thesis, University of Montreal, 195p.

Shannon, C.E., 1949. Communication in the presence of noise. Proc. Institute of Radio Engineers, v. 37, no. 1, 10-21.

Wackernagel, H., 1995. Multivariate Geostatistics. Springer, Berlin.

Yao, T. and Journel, A. G., 1998. Automatic modeling of (cross) covariance tables using fast Fourier transform. Mathematical Geology, v. 30, no. 6, 589-615.

Yaglom, A.M., 1987. Correlation theory of stationary random functions. Springer, New York.

A STATISTICAL TECHNIQUE FOR MODELLING NON-STATIONARY SPATIAL PROCESSES

JOHN STEPHENSON[1], CHRIS HOLMES[2], KERRY GALLAGHER[1] and ALEXANDRE PINTORE[2]
[1] *Dept. Earth Science and Engineering, Imperial College, London.* [2] *Dept. of Mathematics, Oxford University*

Abstract. A deficiency of kriging is the implicit assumption of second-order stationarity. We present a generalisation to kriging by spatially evolving the spectral density function of a stationary kriging model in the frequency domain. The resulting non-stationary covariance functions are of the same form as the evoloved stationary model, and provide an interpretable view of the local effects underlying the process. The method employs a Bayesian formulation with Markov Chain Monte Carlo(MCMC) sampling, and is demonstrated using a 1D Doppler function, and 2D precipitation data from Scotland.

1 Introduction

The standard approach to spatial statistics assumes that the spatial dependence between two points is a function only of separation vector. These procedures fall under the generic label kriging, which are fully described in (Cressie, 1993). Such stationary models however, are unable to take account of localised effects such as geological (e.g. topography, river systems) or political (e.g. state governments conformance to air pollution measures in the US) boundaries, or rapid spatial variations. Although problematic, to date there are few generic non-stationary procedures.

One is the deformation approach of (Sampson and Guttorp, 1992), extended recently to a Bayesian framework in (Damian, Sampson, and Guttorp, 2001) and (Schmidt and O'Hagan, 2003). The more recent kernel-based methods of (Higdon, Swall, and Kern, 1999) and the spectral extension in (Fuentes, 2002) have been shown to be powerful and can be applied when only one observation is available at each site. Other approaches include orthogonal expansion (Nychka and Saltzman, 1998) and the localised moving window approach (Haas, 1990; Haas, 1995). Earlier work is summarised in (Guttorp and Sampson, 1994).

In this paper, we describe and extend our recent generalisation of kriging, involving the spatial evolution of the spectral density function of a stationary process, by manipulation in the frequency domain (Pintore and Holmes, 2003).

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 125-134.

The new method we describe has a variety of attractive aspects, including an interpretable view of the non-stationary process, the definition of a global and analytical covariance structure (thereby making predictions at new locations trivial) and the ability to use the powerful framework developed within kriging directly.

2 Framework for non-stationary covariance functions

Here we use the standard stationary model and evolve a new class of non-stationary process. The emphasis lies in creating new covariance structures that are both non-stationary and interpretable. The proofs for the validity of these theorems can be found in (Pintore and Holmes, 2003).

2.1 STATIONARY GAUSSIAN PROCESSES

In the case of spatial interpolation, we use a stochastic model over the spatial variable s, defined over the p dimensional region $\mathbb{R}^p$. We adopt the standard, stationary approach to spatial statistics and consider our n irregularly sampled data $\mathbf{y}$ to be realisations of a Gaussian process, $Z(s) \sim \mathcal{N}_n(0, \boldsymbol{\sigma^2 \Sigma})$, where $\mathcal{N}_n$ is an n dimensional Gaussian distribution with covariance function Σ, scaled by σ^2. Subsequently we parameterise the covariance function as $\Sigma(s,t) = C(s,t) + \epsilon I_n$, with $C(s,t)$ representing the correlation between two spatial points s and t, ϵ a white noise effect (commonly known as the nugget), and I_n the n dimensional identity matrix. Common forms of $C(s,t)$ include the Gaussian, exponential and Matern stationary correlation functions.

2.2 EVOLUTION IN THE FREQUENCY DOMAIN

The new covariance functions are evolved by modifying stationary covariance functions in the frequency domain. For example, the Gaussian covariance function $C(s,t) = \exp(-\alpha\|s-t\|_p^2)$, has a spectral density function given by $f(\omega) = (4\pi\alpha)^{p/2} exp(\omega'\omega/4\alpha)$, where α is a global smoothing parameter commonly called the range and ω' represents the transpose. Non-stationarity is induced through a localised latent power process $\eta(s)$ acting on the stationary spectrum at location s, hence

$$f_{NS}^{s}(\omega) = h(s)\,[f(\omega)]^{\eta(s)} \tag{1}$$

with the subscript $_{NS}$ now referring to the non-stationary versions of our process, and $h(s)$ a bounded function chosen to ensure constant power in the process. This is in effect saying that when $\eta(s) < 1$, greater emphasis is placed on lower frequencies, producing a smoother process and vice-versa. When $\eta(s) = 1$, we return to the original stationary covariance function. These effects are illustrated in figure 1(a).

We return to the spatial domain via the inverse Fourier transform, producing the non-stationary covariance function C_{NS}. For our example Gaussian function, the final non-stationary covariance function is given by

$$C_{NS}(s,t) = D_{s,t} \exp\left[-\beta_{s,t}\|s-t\|_p^2\right] \tag{2}$$

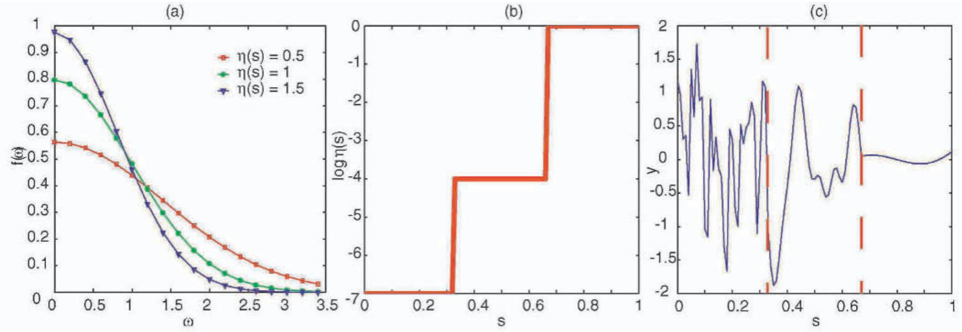


Figure 1. (a) Effect of latent process $\eta(s)$ on the spectral density of a Gaussian covariance function with $\alpha = 0.5$. (b) $\log \eta(s)$ parameterised as a step function. (c) A realisation from a Gaussian covariance function with the latent process defined in (b). The step changeovers are indicated by a dashed line in figure (c).

With

$$\beta_{s,t} = 2\alpha/[\eta(s) + \eta(t)], \tag{3}$$

$$D_{s,t} = 2^{p/2} \frac{[\eta(s)\eta(t)]^{p/4}}{[\eta(s) + \eta(t)]^{p/2}} \tag{4}$$

and is valid for $\eta(s) > 0$. The proof for the validity of this method with respect to Bochner's theorem (see (Levy, 1965)), and the valid choices of $\eta(s)$ are discussed for the Gaussian and Matern covariance functions in (Pintore and Holmes, 2003).

To further illustrate the effect of $\eta(s)$, and possible realisations from such models, we simulate data taken from a Gaussian non-stationary covariance function with $\alpha = 0.5$, and with $\eta(s)$ modelled as a step function (see figures 1(b) and 1(c)). Notice how the realisation $\mathbf{y}$ has high frequency content for $\log \eta(s)$ very negative, and is smooth (low frequency) for $\log \eta(s) \to 0$.

A key point to note is the similarity in form between the stationary and non-stationary covariance functions. This allows us to interpret the latent function $\eta(s)$ directly in terms of the changes in the underlying process.

3 A Bayesian approach

We now present a Bayesian extension to the method, using proper priors. For the moment, we assume a known constant mean of 0 across s to demonstrate the effect of $\eta(s)$ (N.B. The formulation presented is easily extendable to the general case where $Z(s) \sim \mathcal{N}_n(\mathbf{B}\beta, \sigma^2\Sigma)$, with $\mathbf{B}$ representing a matrix of basis functions, and β a scaling vector included to account for deterministic trends in the data). References concerning Bayesian kriging are (Handcock and Stein, 1993) (using reference priors) and (Le and Zidek, 1992).

3.1 LIKELIHOOD, PRIORS AND POSTERIOR DISTRIBUTIONS

For mean equal to 0, the likelihood of the data $\mathbf{y}$ is expressed as

$$p(\mathbf{y}|\theta, \epsilon, \sigma^2) = (2\pi\sigma^2)^{-n/2}|\Sigma|^{-1/2} \exp\left(-\frac{\mathbf{y}'\Sigma^{-1}\mathbf{y}}{2\sigma^2}\right) \tag{5}$$

where Σ is the covariance matrix, parameterised by the nugget effect ϵ and the parameters that define the covariance matrix θ. θ will contain the stationary (α) and the non-stationary (the parameterisation of $\eta(s)$) covariance parameters.

In order to facilitate calculation of the marginals later, we give σ^2 an inverse gamma prior density,

$$p(\sigma^2) \propto (\sigma^2)^{-(a+1)} \exp\left(\frac{-b}{\sigma^2}\right) \tag{6}$$

with the two hyperparameters a and b set to 0.1, providing a wide, non-informative distribution.

For the covariance parameters $p(\theta, \epsilon)$, we assume independent uniform priors, expressing our lack of knowledge about the underlying system, and provide a fully flexible process. As we have assumed independence from σ^2, any other choice of informative prior is equally valid.

These definitions lead to the full posterior, given up to proportionality by Bayes theorem as

$$p(\theta, \epsilon, \sigma^2|\mathbf{y}) \propto (\sigma^2)^{-(n/2+a+1)}|\Sigma|^{-1/2} \exp\left(-\frac{\mathbf{y}'\Sigma^{-1}\mathbf{y} + 2b}{2\sigma^2}\right) p(\theta, \epsilon) \tag{7}$$

3.2 POSTERIOR PREDICTIVE DENSITY

Our goal is to make a prediction of y_0 at a new position in $\mathbb{R}^p$ by integrating the posterior predictive density $p(y_0|\mathbf{y})$. As the integral is intractable, we solve it by sampling from the posterior distribution (equation 7) using MCMC, which for N samples gives the summation,

$$p(y_0|\mathbf{y}) \approx \frac{1}{N}\sum_{i=1}^{N} p(y_0|\theta_i, \epsilon_i, \sigma_i^2) \tag{8}$$

where the first term is the conditional predictive distribution. This density is a shifted t distribution ((Le and Zidek, 1992)) with an expectation (for our simplified case) of $E\left[p(y_0|\theta_i, \epsilon_i, \sigma_i^2)\right] = c_i'\Sigma_i^{-1}\mathbf{y}$, where c_i is the vector of covariates between our data $\mathbf{y}$ and the new data point y_0. Then

$$E\left[p(y_0|\mathbf{y})\right] \approx \frac{1}{N}\sum_{i=1}^{N} c_i'\Sigma_i^{-1}\mathbf{y} \tag{9}$$

in effect generating our average by the summation of the simple kriging predictions for each of the N drawn models i.

3.3 MARKOV CHAIN MONTE CARLO

In order to sample from the posterior distribution, we sample from a Markov chain. The most desirable method is to draw all of the parameters via Gibbs sampling, requiring us to know the full conditional posterior distributions. This however is only possible for the scaling parameter σ^2. After dropping constants in the posterior (equation 7), we have an inverse gamma distribution so that

$$\sigma^2|\mathbf{y}, \theta, \epsilon, \sigma^2 \sim IG(n/2 + a,\ [\mathbf{y}'\Sigma^{-1}\mathbf{y} + 2b]/2) \tag{10}$$

which we can sample from directly. This is not true of the remaining parameters which are tied up in Σ^{-1}, so we use Metropolis-Hastings sampling. To ensure better acceptance rates, we first marginalise σ^2, to give

$$p(\theta, \epsilon|y) = \int (p(\theta, \epsilon, \sigma^2|y)d\sigma^2 \propto p(\theta, \epsilon)|\Sigma|^{-1/2} \left(\frac{\mathbf{y}'\Sigma^{-1}\mathbf{y} + 2b}{2}\right)^{-(n/2+a)} \tag{11}$$

We use Gaussian proposal densities for ϵ and all members of θ, with variances chosen in each case to allow effective traversal of the model space.

4 Results

The two applications chosen to present the properties of our non-stationary method in a Bayesian setting, are a 1D synthetic Doppler function, and a 2D precipitation data set taken from UK Meteorological Office data over Scotland.

4.1 SYNTHETIC DOPPLER FUNCTION

We consider first the Doppler function examined in (Donoho and Johnstone, 1995) and (Pintore and Holmes, 2003), given as

$$f(s) = [s(1-s)]^{1/2} \sin[(2\pi)(1+r)/(s+r)] \quad s \in [0,1] \tag{12}$$

with $r = 0.05$. The sample data $\mathbf{y}$ comprises 128 function evaluations positioned randomly within $s \in [0,1]$ (scaled to have a variance of 7), with added Gaussian white noise (with variance of 1). The function was then predicted at a further 500 points, uniformly sampled in the range $[0,1]$, using the stationary and non-stationary Gaussian covariance functions (equation 2). The accuracy as measured against the true function was then compared. See figure 2(a).

4.1.1 *Latent process formulation*

To parameterise the latent process $\eta(s)$ in the non-stationary covariance function, we follow the suggestion in (Pintore and Holmes, 2003) and use a regression spline with 5 nodes added to a linear function such that

$$\log \eta(s) = \gamma_0 + s\gamma_1 + \sum_{i=1}^{5} \phi_i \|s - u_i\| \tag{13}$$

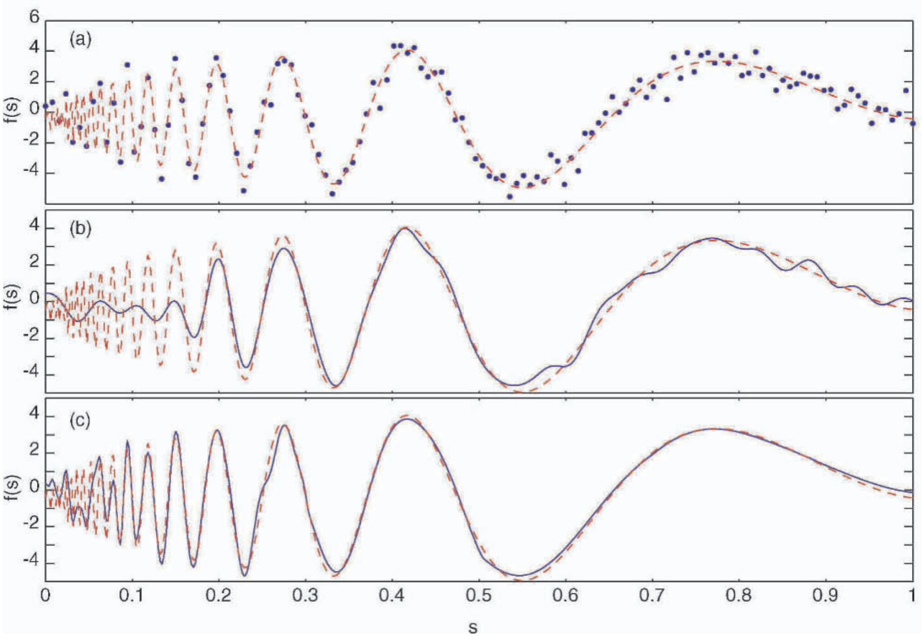


Figure 2. (a) Noisy Doppler function data set, plotted with the dashed predictive data set function. (b) Deterministic fit of the stationary Gaussian covariance function using REML. (c) Posterior predictive fit of the non-stationary Gaussian covariance function (from 4000 samples).

with u_i representing spline knot points, s, and $\{\gamma_0, \gamma_1, \phi\}$ a set of scaling coefficients. In this case, we choose to fix the knot set $\mathbf{u}$, using the kmeans algorithm, and vary the model using only the range, nugget and η scaling coefficients. The covariance parameter vector, θ, now comprises $\{\alpha, \epsilon, \gamma_0, \gamma_1, \phi\}$, which are all sampled using the Metropolis Hastings algorithm. From a run of 5000 iterations, the first 1000 were discarded as 'burn-in', and the remaining used to find the mean posterior prediction.

The stationary model was fitted using the technique of restricted maximum likelihood (REML) (Cressie, 1993), and optimised using a deterministic search (Nelder Mead algorithm).

4.1.2 *Prediction Comparison*

A comparison of the predictive powers of both samples can be found in figures 2(b) and 2(c). It is evident that the non-stationary method has performed far better than in the stationary case, which has been forced to comprimise between the very smooth data as $s \to 1$, and the higher frequencies as $s \to 0$.

In figure 3 we give the mean and 95% confidence intervals (an immediate advantage of using MCMC sampling) over the posterior latent process $\eta(s)$ for the 4000 samples. The figure is interpreted as placing higher emphasis on lower frequencies where $s < 0.24$ (where $\log \eta(s) < 0$), whilst providing an increasingly smooth covariance as $s \to 1$. This is further explained via figure 1(a).

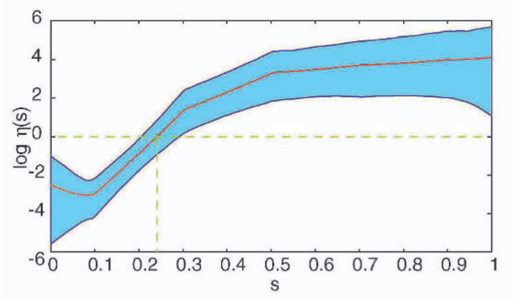


Figure 3. Log of the mean and 95% confidence intervals for the latent process $\eta(s)$. The stationary case corresponds to $\log \eta(s) = 0$.

4.2 PRECIPITATION IN SCOTLAND

For a real data example, we consider the UK Met Office 'Land Surface Observation Stations Data' held at the British Atmospheric Data Centre (BADC, 2004). The analysis of precipitation is important for agricultural as well as environmental reasons (eg. pollutant dispersal following the Chernobyl disaster). The purpose of this analysis was to demonstrate the interpretability of the latent process, rather than a comparison of stationary to non-stationary covariance functions.

Specifically, we extracted daily rainfall measurements from 1997, for the months of January, February and March, and analysed the measurements for 481 land stations in Scotland. The daily measurements within these three months were then averaged, to give three distinct data sets of average daily rainfall, (millimetres per day). A scatter plot of the January data is shown in figure 4.

4.2.1 *Latent process formulation*

To parameterise $\eta(\mathbf{s})$ (where $\mathbf{s}$ is now a 2 dimensional vector $[s_1, s_2]$ corresponding to longitude and latitude respectively), we choose to use the thin-plate spline with the form

$$\log \eta(\mathbf{s}) = \gamma_0 + \gamma_1 s_1 + \gamma_2 s_2 + \sum_{i=1}^{k} \phi_i \|\mathbf{s} - \mathbf{u}_i\|_p^2 \ \log \|\mathbf{s} - \mathbf{u}_i\|_p^2 \qquad (14)$$

where $\mathbf{u}_i$ is a set of k knot points, and $\{ \gamma_0, \gamma_1, \gamma_2\}$ the scaling coefficients. In this case, we take $k = 20$ and fix the positions of the knot points using the kmeans clustering algorithm. Again choosing the Gaussian non-stationary covariance function, we perform MCMC over the spline scaling coefficients and the stationary parameters ϵ and α, again discarding the first 1000 iterations as 'burn-in', and averaging over the remaining 4000 samples.

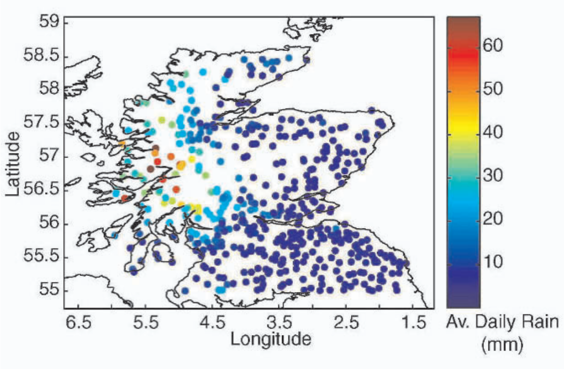


Figure 4. Positions and values of the average daily rainfall for Scotland during January 1997

4.2.2 *Interpretation of* $\eta(s)$ *in January*

We first look at the non-stationarity revealed in the January data set. A contour plot of the mean of $\eta(\mathbf{s})$, (see figure 5(a)) reveals a strong trend, moving from low values of $\eta(\mathbf{s})$ in the west, to high values in the east. Thus on the west coast, where rain is far more prevalent and protection from other landmasses is minimal, there is much greater variation from station to station. This is demonstrated by the greater emphasis placed on higher frequencies by $\eta(s)$, giving a local covariance function with a small range of influence. This contrasts with the smoother region in the east, where the level of rainfall is reduced, sheltered as it is by the topography in the Highlands.

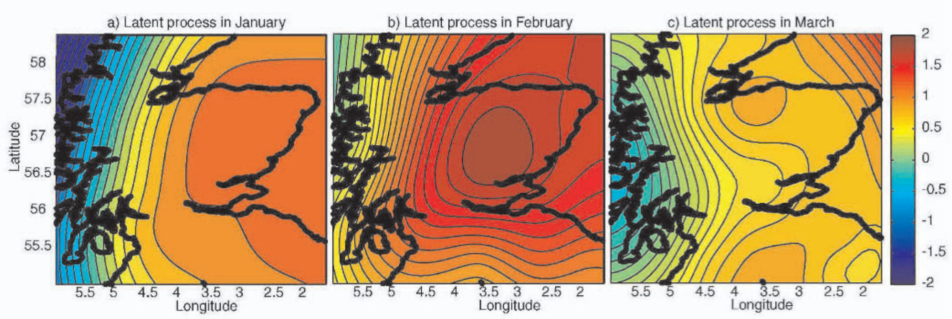

a) Latent process in January
b) Latent process in February
c) Latent process in March
Latitude
58
57.5
57
56.5
56
55.5
5.5
5
4.5
4
3.5
3
2.5
2
Longitude
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2

Figure 5. Contour plots of the latent process for (a) January, (b) February and (c) March during 1997.

To illustrate the significance of this west-east trend, we take a projection in the longitudinal direction and plot the 95% credible intervals (figure 6(a)). This demonstrates the small variation in the north-south direction, and the relatively tight credible intervals, reinforcing the notion of an west-east trend in stationarity.

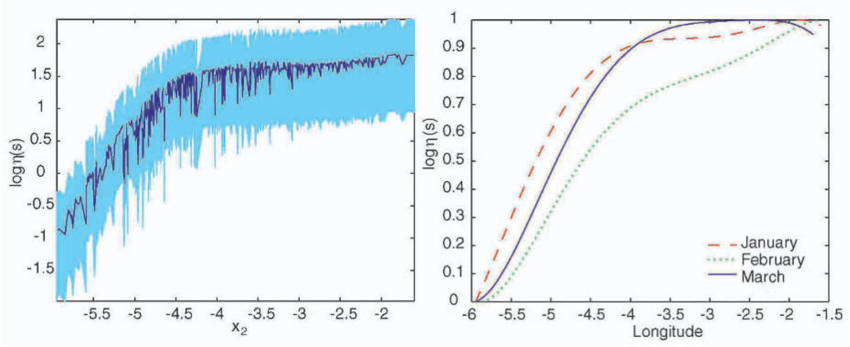


Figure 6. (a) log of the mean of $\eta(s)$ projected onto the longitudinal direction, plotted with 95% credible intervals. (b) Comparison of scaled longitudinal projections of $\eta(s)$ for the January, February and March data sets.

4.2.3 *Comparison of $\eta(s)$ for three different months*

To further demonstrate the importance of this east west trend, we carried out the same analysis on the two subsequent months and compared the posterior values of $\eta(s)$. These data sets are compared directly in figure 5, comprising a contour plot for each of the three months and again show strong west-east trends.

As the absolute values of $\eta(s)$ in these figures are influenced by the effect of the stationary covariance parameter α, we compare the longitudinal projections of $\log \eta(s)$ by first fitting a polynomial and then scaling the values to the range [0, 1]. This is shown in figure 6(b), and demonstrates concisely the consistency of the recognised trend.

The consistency of this result indicates that there is an underlying process causing a consistent non-stationarity in the data. Suggestions as to the cause of this observation are geographical effects such as coastal regions, shielding by topography and the consistent direction of weather patterns. Significantly, this demonstrates the ability of the method to reveal an accurate measure of the non-stationarity from only one realisation.

5 Discussion

In summary, we have presented a method for inducing non-stationarity in standard covariance functions, by use of latent power process $\eta(s)$ in the frequency domain. Such a treatment yields covariance functions that have the same analytical form of the base stationary function, and offers a direct and interpretable view of any non-stationary processes. A Bayesian methodology has been used and demonstrated using MCMC providing access to the full uncertainties.

The two applications have revealed an increased prediction accuracy when compared to standard stationary techniques, and demonstrated the ability to extract the underlying non-stationarity from a single realisation.

Now that the Bayesian context has been established, future work will involve using reversible jump MCMC when inferring $\eta(s)$ (Green, 1995), (providing the ability to change the position and number of knot points), as well as applying the method to spatio-temporal processes by treating η as a function of space and time.

Acknowledgements

We are grateful to the British Atmospheric Data Centre for providing access to the Met Office Land Surface Observation Stations Data, and to the UK Natural Environment Research Council for funding this work.

References

BADC, *Met Office - Land Surface Observation Stations Data,* http://badc.nerc.ac.uk/data/surface/, 2004.

Cressie, N. A. C., *Statistics for spatial data,* Wiley series in probability and mathematical statistics. Applied probability and statistics., 1993.

Damian, D., Sampson, P. D. and Guttorp, P., *Bayesian estimation of semi-parametric non-stationary spatial covariance structures,* Environmetrics, vol. 12, no. 2, 2001, p. 161-178.

Donoho, D. L. and Johnstone, I. M., *Adapting to unknown smoothness via wavelet shrinkage,* Journal of the American Statistical Association, vol. 90, no. 432, 1995, p. 1200-1224.

Fuentes, M., *Spectral methods for nonstationary spatial processes,* Biometrika, vol. 89, no. 1, 2002, p. 197-210.

Green, P. J., *Reversible jump Markov chain Monte Carlo computation and Bayesian model determination,* Biometrika, vol. 82, no. 4, 1995, p. 711-732.

Guttorp, P. and Sampson, P. D., *Methods for estimating heterogeneous spatial covariance functions with environmental applications,* in G.P. Patil and C.R. Rao (eds), Handbook of Statistics XII: Environmental Statistics. Elsevier, 1994, p. 663-90.

Haas, T. C., *Lognormal and Moving Window Methods of Estimating Acid Deposition,* Journal of the American Statistical Association, vol. 85, no. 412, 1990, p. 950-963.

Haas, T. C., *Local prediction of a spatio-temporal process with an application to wet sulfate deposition,* Journal of the American Statistical Association, vol. 90, no. 432, 1995, p. 1189-1199.

Handcock, M. S. and Stein, M. L., *A Bayesian-Analysis of Kriging,* Technometrics, vol. 35, no. 4, 1993, p. 403-410.

Higdon, D., Swall, J. and Kern, J., *Non-stationary spatial modeling,* in Bernado, J.M. et al. (eds), *Bayesian statistics 6 : proceedings of the Sixth Valencia International Meeting, June 6-10, 1998.* Oxford University Press, 1999.

Le, N. D. and Zidek, J. V., *Interpolation with Uncertain Spatial Covariances - a Bayesian Alternative to Kriging,* Journal of Multivariate Analysis, vol. 43, no. 2, 1992, p. 351-374.

Levy, P., *Processus stochastiques et mouvement brownien.* Gauthier-Villars, 1965.

Nychka, D. and Saltzman, N., *Design of Air Quality Monitoring Networks,* Case Studies in Environmental Statistics, no. 132, 1998, p. 51-76.

Pintore, A. and Holmes, C. C., *Constructing localized non-stationary covariance functions through the frequency domain,* Technical Report. Imperial College, London, 2003.

Sampson, P. D. and Guttorp, P., *Nonparametric-Estimation of Nonstationary Spatial Covariance Structure,* Journal of the American Statistical Association, vol. 87, no. 417, 1992, p. 108-119.

Schmidt, A. M. and O'Hagan, A., *Bayesian inference for non-stationary spatial covariance structure via spatial deformations,* Journal of the Royal Statistical Society Series B-Statistical Methodology, vol. 65, no. 3, 2003, p. 743-758.

CONDITIONING EVENT-BASED FLUVIAL MODELS

MICHAEL J. PYRCZ and CLAYTON V. DEUTSCH
Department of Civil & Environmental Engineering, 220 CEB,
University of Alberta, Canada, T6G 2G7

Abstract. A fluvial depositional unit is characterized by a central axis, denoted as a streamline. A set of streamlines can be used to describe a stratigraphic interval. This event-based (denoted as event-based to avoid confusion with streamline-based flow simulation) approach may be applied to construct stochastic fluvial models for a variety of reservoir types, fluvial styles and systems tracts. Prior models are calculated based on all available soft information and then updated efficiently to honor hard well data.

1 Introduction

Interest in North Sea fluvial reservoirs led to the development of object-based models for fluvial facies and geometries (see Deutsch and Wang, 1996 for a review of development). For these models conditioning is often problematic. These difficulties in conditioning spurred research in direct object modeling. Visuer et al. (1998) and Shmaryan and Deutsch (1999) published methods to simulate fluvial object-based models that directly honor well data. These algorithms segment the well data into unique channel and nonchannel facies and then fit channels through the segments. The channel center line is parameterized as a random function of departure along a vector and the geometry is based on a set of sections fit along the center line.

Yet, these techniques are only well suited to paleo valley (PV) reservoir types. The PV reservoir type geologic model is based on ribbon sandbodies from typically low net-to-gross systems with primary reservoir quality encountered in sinuous to straight channels and secondary reservoir rock based on levees and crevasse splays embedded in overbank fines (Galloway and Hobday, 1996; Miall, 1996).

More complicated channel belt (CB) fluvial reservoir types are common. Important examples include the McMurray Formation (Mossop and Flach, 1983, Thomas et al., 1987) and Daqing Oil Field, China (Jin et al., 1985, Thomas et al., 1987). These reservoirs include complicated architectural element configurations developed during meander migration punctuated by avulsion events. The application of the bank retreat model for realistic channel meander migration has been proposed by Howard (1992), applied by Sun et al. (1996) and Lopez et al. (2001)

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 135-144.

to construct realistic models of CB type fluvial reservoirs. These methods lack flexibility in conditioning.

A event-based paradigm is introduced with (**1**) improved flexibility to reproduce a variety of fluvial reservoir styles with realistic channel morphologies, avulsion and meander migration and (**2**) a new efficient approach to condition to well data and areal reservoir quality trends. Fortran algorithms are available that apply this techniques, `ALLUVSIM` is an unconditional algorithm for the construction of training images and the `ALLUVSIMCOND` algorithm includes streamline updating for well conditioning. Greater detail on this work and the associated code is available in Pyrcz (2004).

This work was inspired by the developments of Sun et al. (1996) and Lopez et al. (2001), but it was conducted independent of Cojan and Lopez (2003) and Cojan et al. (2004). The reader is referred to these recent papers for additional insights into the construction of geostatistical fluvial models.

2 Event-based Stochastic Fluvial Model

The basic building block of this model is the *streamline*. A streamline represents the central axis of a flow event and backbone for architectural elements (Wietzerbin and Mallet, 1993). This concept is general and may represent confined or unconfined, fluvial or debris flows.

Genetically related streamlines may be grouped into *streamline associations*. Streamline associations are interrelated by process. For example, a streamline association may represent a channel fill architectural elements within a braided stream or lateral accretion architectural elements within point bar. Fluvial architectural elements are attached to streamlines and architectural element interrelationships are characterized by streamline associations. This is a logical technique for constructing fluvial models since all architectural elements are related to “flow events”.

2.1 3-D STREAMLINES

The direct application of a cubic spline function to represent the plan view projection of a fluvial flow event is severely limited. As a function, a spline represented as $f^s(x)$ may only have a single value for any value x. In graphical terms, a function may not curve back on itself. This precludes the direct use of a spline function to characterize high sinuosity channel streamlines.

A streamline is modeled as a set of cubic splines. Each spline models the coordinates (x, y and z) with respect to distance along the spline (s). The advantages of this technique are: (**1**) continuous interpolation of streamline location in Cartesian coordinates at any location along the streamline, (**2**) relatively few parameters required to describe complicated curvilinear paths, (**3**) manipulation of splines is much more computationally efficient than modifying geometries and (**4**) other properties such as architectural element geometric parameters and longitudinal trends may be stored as continuous functions along the streamline. These issues are discussed in further detail below.

The control nodes of a 3-D spline may be freely translated. The only requirement is that the second derivatives of the spline location parameters is recalculated after modification. This operation is very fast. The calculation of complicated geometries generally requires a high level of computational intensity or simplification. In the event-based models the geometric construction is postponed to the end of the algorithm. This results in very fast calculation and manipulation of complicated geometric morphologies and associations represented as 3-D splines.

Any properties may be attached to the 3-D spline and interpolated along the length of the spline. In the fluvial event-based model, the channel width, local curvature, relative thalweg location and local azimuth are included in the 3-D spline. Other information including architectural element type and additional property trends may be included. These properties are calculated at the control nodes and then splines are fit as with the location parameters.

2.2 STREAMLINE ASSOCIATIONS WITHIN EVENT-BASED MODELS

A streamline association is a grouping of interrelated 3-D splines. Streamline associations are characterized by their internal structure and interrelationship or stacking patterns. The internal structure is the relation of streamlines within the streamline association. The external structure is the interrelationship between streamline associations. Streamline associations may be tailored to reproduced features observed in each fluvial reservoir style.

A variety of stacking patterns may exist in the fluvial depositional setting. Compensation is common in dispersive sedimentary environments such as proximal alluvial fans, vertical stacking with little migration is common in anastomosing reaches and nested channel belts often form in incised valleys. These patterns include important information with regard to the heterogeneity of a reservoir and should be included in fluvial models.

2.3 STREAMLINE OPERATIONS

A suite of streamline operations is presented that allow for event-based models to be constructed by the creation and modification of streamlines. These operations include (**1**) initialization, (**2**) avulsion, (**3**) aggradation and (**4**) migration.

The streamline *initialization operator* is applied to generate an initial streamline or to represent channel avulsion proximal of the model area. The disturbed dampened harmonic model developed by Ferguson (1976) is applied.

The *avulsion operator* creates a copy of a specific channel streamline, selects a location along the streamline, generates a new downstream channel segment with same streamline sinuosity and the same geometric parameter distributions. The geometric parameters (e.g. channel width) of the new streamline are corrected so that the properties are continuous at the avulsion location. The initial azimuth is specified as the azimuth of the tangent at the avulsion location. There is no constraint to prevent the avulsed streamline from crossing the original streamline distal of the avulsion location.

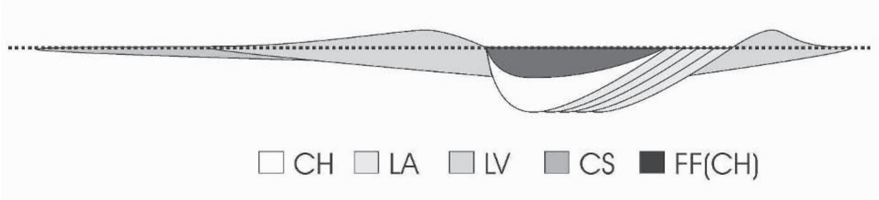


Figure 1. An illustration of the fluvial architectural elements applied in the event-based model.

Aggradation is represented by a incremental increase in the elevation of a streamline. The current implementation is to add a specified constant value to the elevation, z, parameter for all control nodes.

The streamline *migration operator* is based on the bank retreat model. The application of the bank retreat model for realistic channel meander migration has been proposed by Howard (1992), applied to construct fluvial models by Sun et al. (1996) and extended to construct meandering fluvial models that approximately honor global proportions, vertical and horizontal trends by Lopez et al. (2001).

Key implementation differences from the original work from Sun et al. (1996) include (**1**) standardization of migration steps, (**2**) integration of 3-D splines for location and properties, (**3**) application of various architectural elements. The meander migration along the streamline is standardized such that the maximum migration matches a user specified value. This removes the significance of hydraulic parameters such as friction coefficient, scour factor and average flow rate, since only the relative near bank velocity along the streamline is significant. Hydraulic parameters are replaced by the maximum spacing of accretion surfaces, which may be more accessible in practice.

2.4 FLUVIAL ARCHITECTURAL ELEMENTS

The available architectural elements include (**1**) channel fill (CH), (**2**) lateral accretion (LA), (**3**) levee (LV), (**4**) crevasse splay (CS), (**5**) abandoned channel fill (FF(CH)) and (**6**) overbank fines (FF) (see illustration in Figure 1). The geometries and associated parameters are discussed for each element in detail in Pyrcz (2004).

2.5 EVENT SCHEDULE

The event-based approach is able to reproduce a wide variety of reservoir styles with limited parametrization. This algorithm may reproduce braided, avulsing, meandering channels and may reproduce geometries and interrelationships of a variety of fluvial reservoir types. The algorithm is supplied with areal and vertical trends, distributions of geometric parameters, probabilities of events and architectural elements.

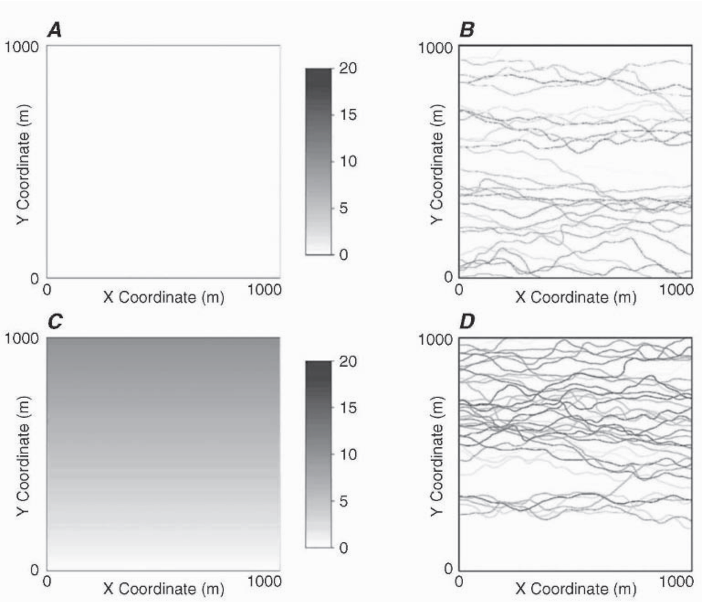


Figure 2. Example areal trends in channel density and the resulting streamlines. A and B - no areal trend supplied and C and D - a linear trend increasing in the y positive direction. Note areal trend is a relative measure without units.

2.6 AREAL CHANNEL DENSITY TRENDS

Analogue, well test and seismic information may indicate areal trends in reservoir quality. Although seismic vertical resolution is often greater than the reservoir thickness, seismic attributes calibrated to well data may indicate a relative measure of local reservoir quality. Well tests may provide areal information on the distribution of reservoir quality and may significantly constrain model uncertainty. Analogue information such as reservoir type may indicate a confined PV type or a more extensive and uniform SH type reservoirs. If the net facies are associated with CH, LV and CS elements then this areal trend information may be integrated by preferentially placing streamlines in areal locations with high reservoir quality.

The technique for honoring areal trends is to (**1**) construct a suite of candidate streamlines with the desired morphology, (**2**) superimpose each candidate streamline on the areal trend model and calculated average relative quality along the streamline and (**3**) for each streamline initialization drawn from this distribution of candidate streamlines (without replacement) weighted by the average quality index. This technique is efficient since the construction of hundreds or thousands of streamlines is computationally fast. This technique is demonstrated in Figure 2.

2.7 VERTICAL CHANNEL DENSITY TRENDS AND AGGRADATION SCHEDULE

Well data and analogue information may provide information on vertical trends in reservoir quality. Well logs calibrated by core are valuable sources of vertical trend information. Often, identification of systems tract and fluvial style will provide analogue information concerning potential vertical trends.

These trends may be honored by constraining the aggradation schedule. The current implementation is to apply the trend within a user defined number of constant elevation levels. Streamlines and associated architectural elements are generated at the lowest level until the NTG indicated by the vertical trend is reached for the model subset from the base of the model, to the elevation of the first level. Then the aggradation operator is applied to aggrade to the next level and the process is repeated through all user defined levels. For the highest level, the model is complete when the global NTG ratio is reached.

3 Conditional Event-based Simulation

There are a variety of available methods that may be applied to condition complicated geologic models; (**1**) dynamically constrain model parameters during model construction to improve data match (Lopez et al., 2001), (**2**) posteriori correction with kriging for conditioning (Ren et al.,2004), (**3**) pseudo-reverse modeling (Tetzlaff, 1990), (**4**) apply as a training image for multiple-point geostatistics (Strebelle, 2002) and (**5**) direct fitting of geometries to data (Shmaryan et. al., 1999 and Visuer et. al., 1998). Each of these techniques has limitations either in efficiency, robustness or the ability to retain complicated geometries and interrelationships.

An event-based model consists of associations of streamlines with associated geometric parameters and identified architectural elements. A prior model of streamline associations may be updated to reproduce well observations. The proposed procedure is: (**1**) construct the prior event-based model conditioned by all available soft information, (**2**) interpret well data and identify CH′ element intervals (where CH′ elements are channel fill elements without differentiation of CH, LA and FF(CH) elements), (**3**) update streamline associations to honor identified CH′ element intervals and (**4**) correct for unwarranted CH′ intercepts. This technique entails the manipulation of large-scale elements to honor small scale data; therefore, it is only suitable for settings with sparse conditioning data. Settings with dense data may be intractable.

3.1 INTERPRETED WELL DATA

The hard data from wells is applied to identified CH′ element intervals. CH′ elements are typically identified by erosional bases and normal grading. CH′ element fills often occur in multistory and multilateral configurations. CH′ elements often erode into previously deposited CH′ elements to form amalgamated elements (Collinson, 1996, Miall, 1996).

The geologic interpretation of well data is performed prior to the updating step. The input data includes the areal location for each vertical well and a list of CH′ element intervals with base and original top (prior to erosion). The geologic interpretation is often uncertain, especially with amalgamated CH′ elements. Alternate geologic interpretations may be applied to account for this uncertainty.

3.2 UPDATING STREAMLINE ASSOCIATIONS TO HONOR WELL DATA

The model is updated by modifying the position of streamline associations to honor CH′ element intercepts observed in well data. For each CH′ element interval the following steps are performed. (**1**) The horizontal position is corrected such that the CH′ element intercept thickness is within tolerance of the CH′ element interval thickness. (**2**) Then the vertical location is corrected such that the CH′ element intercept top matches the top of the CH′ element interval. Entire streamline associations are corrected to preserve the relationships between streamlines within a streamline association. For example, if a streamline association includes a set of streamlines related by meander migration, the entire set of streamlines representing a point bar is shifted. If individual streamlines were modified independently this may change the nature of the streamline association.

The CH′ element intervals are sequentially corrected. If there is no previous conditioning then streamline associations are translated (see A in Figure 3). If there is previous conditioning a smooth correction method is applied to the streamline association (see B in Figure 3). A step vector is constructed oriented from the nearest location on a streamline within the streamline association to the location of the well interval. The scale of the step of the sense is determined by an iterative procedure described below.

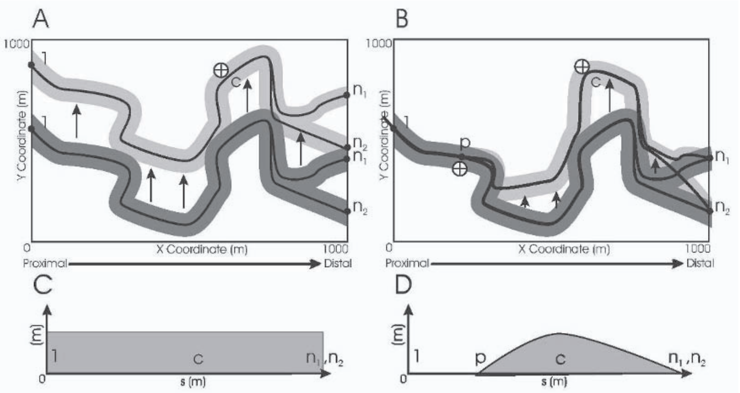


Figure 3. An illustration of methods for updating streamline associations with well data. For this example, there are two streamlines in the streamline association representing an avulsion event that are corrected to honor conditioning data (c). A - the case with no previous conditioning. B - the case with previous conditioning. C and D - the transverse correction with respect to location along the streamline.

3.3 ITERATIVE PROCEDURE FOR UPDATING STREAMLINE ASSOCIATIONS

Modifications of streamline associations has an impact on CH′ element geometry. It would be difficult to directly calculate the precise translation of a streamline to result in the correct interval thickness at a well location. A simple iterative method is applied to correct the well intercept thickness. The thickness of the CH′ element from a streamline association is calculated at the vertical well location. The error is calculated, if the thickness is less than indicated by the conditioning then the streamline association is shifted towards the well location. If the thickness is greater than indicated by the conditioning then the streamline association is shifted away from the well location. The procedure is repeated for all identified CH′ element intercepts.

3.4 CORRECTION FOR UNWARRANTED WELL INTERCEPTS

The correction for unwarranted CH′ element intercepts applies a robust iterative technique. For each unwarranted CH′ element intercept the associated streamline association is checked for conditioning. If the streamline association is not anchored to conditioning data then the streamline association may be translated in the direction transverse to the primary flow direction. If the streamline association is anchored to conditioning data then a smooth modifications is applied.

The streamline association is modified until the thickness of the unwarranted CH′ element intercept reaches zero. For each iteration the step size of the modification is increased and the direction is reversed. This method is robust since it does not become trapped with complicated streamline associations. This methodology is illustrated in Figure 4 with a complicated setting.

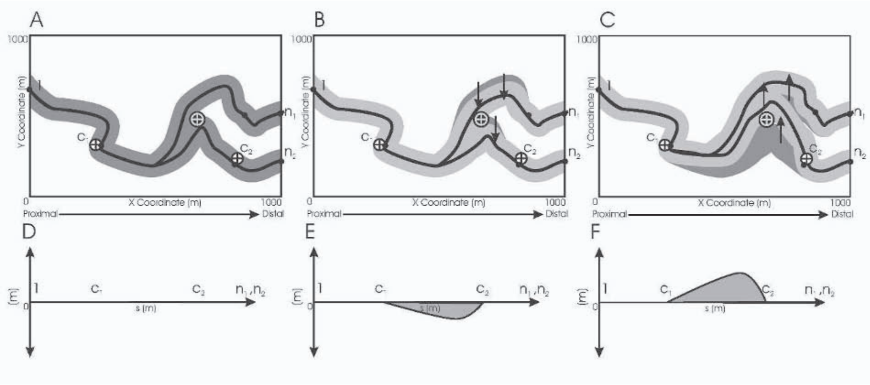


Figure 4. An illustration of the method for correcting streamline associations to remove unwarranted well intercepts. The two streamlines are related by avulsion in the streamline association and there are two previously conditioned locations (C_1 and C_2). A and D - the initial streamline association prior to correction. B and E - the first smooth modification (Oliver, 2002). C and F - the second iteration.

3.5 EXAMPLE CONDITIONAL EVENT-BASED MODELS

The `ALLUVSIMCOND` algorithm was applied to construct a conditional model. The streamlines include braided low to high sinuosity morphology. A single well is included with two CH′ element intervals identified. Cross sections and streamline plan sections of the prior and updated models are shown in Figure 5. The morphology of the streamlines is preserved while the well intercepts are honored.

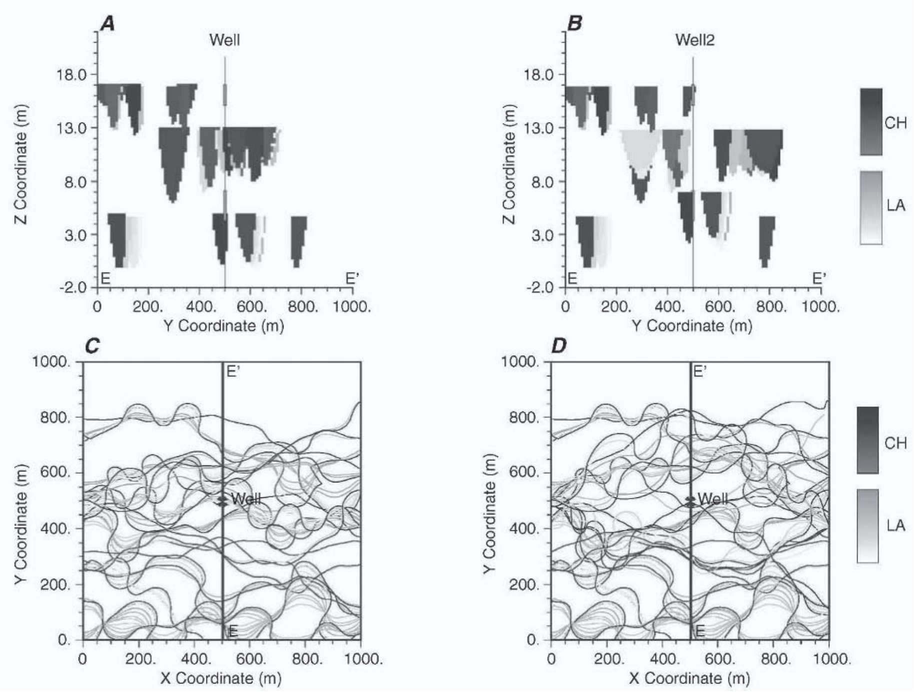


Figure 5. An example conditional event-based model from ALLUVSIMCOND. A and B - cross section of prior and updated model and C - D plan section of prior and updated model streamlines with cross section indicated.

4 Conclusions and Future Work

The event-based approach is a flexible and efficient tool for the construction of stochastic fluvial models. The building block approach allows for the modeling of a variety of fluvial reservoir styles, including the complicated architectures of CB type fluvial reservoirs. Event-based models may be constructed based on all available soft geologic information and then updated to honor hard well data.

Future implementation will address well observations of other architectural elements and the applications of the the event-based approach to a variety of depositional settings, such as deepwater (Pyrcz, 2004).

Acknowledgements

We would like to acknowledge the industry sponsors of the Centre for Computational Geostatistics at the University of Alberta and the National Science and Engineering Council of Canada for supporting this research.

References

Cojan, I. and Lopez, S., *Process-based Stochastic Modeling of Meandering Channelized Reservoirs*, AAPG Annual Meeting, Dallas, 2003, May

Cojan, I. and Fouche, O and Lopez, S., *Process-based Reservoir Modelling in the Example Meandering Channel*, Seventh International Geostatistics Congress, Banff, 2004 September

Collinson, J. D., *Alluvial Sediments*, *in* H. G. Reading, editor, *Sedimentary Environments: Processes, Facies and Stratigraphy*, Balckwell Science, 1996.

Deutsch, C. V. and Tran, T. T., *FLUVSIM: A Program for Object-Based Stochastic Modeling of Fluvial Depositional Systems*, Computers & Geosciences, vol. 28, 2002.

Ferguson, R. I., *Disturbed Periodic Model for River Meanders*, Earth Surface Processes, vol. 1, p. 337-347.

Galloway, W. E. and Hobday, D. K., *Terrigenous Clastic Depositional Systems*, Springer, 1997.

Howard, A. D., *Modeling Channel Migration and Floodplain Sedimentation in Meandering Streams*, *in* P. A. Carling and G. E. Petts, editor, *Lowland Floodplain Rivers*, John Wiley and Sons, 1992.

Jin, Y., Liu D. and Luo C., *Development of Daqing Oil Field by Waterflooding*, Journel Petroleum Technology, February, 1985, p. 269-274.

Lopez, S, Galli, A. and Cojan, I., *Fluvial Meandering Channelized Reservoirs: a Stochastic and Process-base Approach*, 2001 Annual Conference of the IAMG, International Association of Mathematical Geologists, Cancun, Mexico, 2001.

Miall, A. D., *The Geology of Fluvial Deposits*, Springer, 1996.

Mossop, G. D. and Flach, P. D., *Deep Channel Sedimentation in the Lower Cretaceous McMurray Formation*, Sedimentology, vol. 30, 1983, p. 493-509.

Oliver, D. S., *Conditioning Channel Meanders to Well Observations*, Mathematical Geology, vol. 34, 2002, p. 185-201.

Pyrcz, M. J. *Integration of Geologic Information into Geostatistical Models*, Ph. D. Thesis, University of Alberta, Edmonton, 2004.

Ren, W., Cunha, L. and Deutsch, C. V., *Preservation of Multiple Point Structure when Conditioning by Kriging*, Sixth International Geostatistics Congress, Banff, AB, 2004

Shmaryan, L. and Deutsch, C. V., *Object-Based Modeling of Fluvial/Deepwater Reservoirs with Fast Data Conditioning: Methodology and Case Studies*, SPE Annual Technical Conference and Exhibition, Society of Petroleum Engineers, Houston, TX, 1999.

Strebelle, S., *Conditional Simulation of Complex Geological Structures Using Multiple-Point Statistics*, Mathematical Geology, vol. 32, no. 9, 2002, p. 2937-2954.

Sun T., Meakin, P. and Josang, T., *A Simulation Model for Meandering Rivers*, Water Resources Research, vol. 32, no. 9, 1996, p. 2937-2954.

Tetzlaff, D. M., *Limits to the Predictive Ability of Dynamic Models the Simulation Clastic Sedimentation*, *in* T. A. Cross, editor, *Quantitative Dynamic Stratigraphy*, Prentice-Hall, 1990.

Thomas, R. G., Smith D. G., Wood, J. M., Visser, J., Calverley-Range, A. and Koster, E. H., *Inclined Heterolithic Stratification - Terminology, Description, Interpretation and Significance*, Sedimentary Geology, vol. 53, 1987, p. 123-179.

Viseur, S., Shtuka, A. and Mallet J. L., *New Fast, Stochastic, Boolean Simulation of Fluvial Deposits*, SPE Annual Technical Conference and Exhibition, Society of Petroleum Engineers, New Orleans, LA, 1998.

Wietzerbin, L. J. and Mallet J. L., *Parameterization of Complex 3D Heterogeneities: A New CAD Approach*, SPE Annual Technical Conference and Exhibition, Society of Petroleum Engineers, Houston, TX, 1993.

3D GEOLOGICAL MODELLING AND UNCERTAINTY: THE POTENTIAL-FIELD METHOD

CHRISTOPHE AUG [(1)], JEAN-PAUL CHILÈS [(1)],
GABRIEL COURRIOUX [(2)] and CHRISTIAN LAJAUNIE [(1)]
[(1)] Centre de Géostatistique, Ecole des Mines de Paris
35, rue Saint-Honoré
77305 Fontainebleau Cedex, France
[(2)] Bureau de Recherches Géologiques et Minières
3, avenue C. Guillemin – BP 6009
45062 Orléans Cedex 2, France

Abstract. The potential-field method (Lajaunie *et al.*, 1997) is used to create geological surfaces by interpolating from points on interfaces, orientation and fault data by universal cokriging. Due to the difficulty of directly inferring the covariance of the potential field, it is identified from the orientation data, which can be considered as derivatives of the potential. This makes it possible to associate sensible cokriging standard deviations to the potential-field estimates and to translate them into uncertainties in the 3D model.

1 Introduction

During the last ten years, 3D geological modelling has become a priority in several domains such as reservoir characterization or civil engineering. In geological mapping too, 3D digital pictures are created to model and visualize the subsurface and the relations between layers, faults, intrusive bodies, etc. While completing its 1:50 000 geological map programme for the entire French territory, B.R.G.M. (the French geological survey) started a research project for defining three-dimensional maps which could clearly represent the subsurface and underground geology. A new tool, the "Editeur Géologique", has been developed to face this particularly tough issue. It is based on the construction of implicit surfaces using the potential-field method.

2 Reminders on the potential-field method

2.1 PRINCIPLES

The problem is to model the geometry of geological layers using drill-hole data, digital geological maps, structural data, interpreted cross-sections, etc.
The method is based on the interpolation of a scalar field considered as a potential field. In this approach, a surface is designed as a particular isovalue of the field in 3D space.
In all the following equations, **x**=(x,y,z) is a point in the three-dimensional space R^3. The potential is assumed to be a realization of a differentiable random function Z.

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 145-154.

We first consider one interface or several sub-parallel interfaces (iso-surfaces related to the same potential field) and we will see later how to manage several fields.

2.2 DATA

The 3D model is obtained by integrating data originating from different sources.
The first kind of data is a set of points belonging to the interfaces to be modelled. They come from digitized contours on the geological map and from intersections with boreholes. The other type of data is structural data (orientation of surfaces).
For the interpolation of the potential field, these data are coded as follows:
- if we have a set J of n points on an interface, we use n-1 linearly independent increments of potential, all equal to zero; these increments are of the form:

$$\begin{aligned} &Z(\mathbf{x_j}) - Z(\mathbf{x_{j'}}) = 0 \\ &e.g., \quad Z(\mathbf{x_j}) - Z(\mathbf{x_{j-1}}) = 0 \quad j = 2,\dots,n \end{aligned}$$

If several interfaces are modelled with the same potential field, the data set J is the union of the elementary data sets relative to the various interfaces.
- orientation data are considered as gradients of the potential, namely polarized unit vectors, normal to the structural planes:

$$\frac{\partial Z}{\partial x}(\mathbf{x_i}) = G_i^x, \quad \frac{\partial Z}{\partial y}(\mathbf{x_i}) = G_i^y, \quad \frac{\partial Z}{\partial z}(\mathbf{x_i}) = G_i^z$$

2.3 SOLUTION

Determining a geological interface is an interpolation problem which can be solved by determining the potential at any point in the space and by drawing the iso-potential surface corresponding to the interface. The potential field is defined up to an arbitrary constant, because we only work with increments. Indeed we will interpolate the potential at $\mathbf{x}$ in comparison with the potential at some reference point $\mathbf{x_0}$. These increments of potential are estimated as:

$$\left[Z(\mathbf{x}) - Z(\mathbf{x_0})\right]^* = \sum_i \left(\lambda_i G_i^x + \mu_i G_i^y + \nu_i G_i^z\right) + \sum_j \lambda_j \left[Z(\mathbf{x_j}) - Z(\mathbf{x}_{j-1})\right] \quad (1)$$

The last term is equal to zero, but we introduce it here, because the weights λ_i, μ_i, ν_i are different from weights based on the gradient data alone.
The weights are the solution of a universal cokriging system of the form:

$$\begin{pmatrix} C_G & {}^tC_{GI} & {}^tU_G & {}^tF_G \\ C_{GI} & C_I & {}^tU_I & {}^tF_I \\ U_G & U_I & 0 & 0 \\ F_G & F_I & 0 & 0 \end{pmatrix} \begin{pmatrix} A \\ B \\ C \\ D \end{pmatrix} = \begin{pmatrix} C_G^0 \\ C_I^0 \\ U^0 \\ F^0 \end{pmatrix}$$

C_G and C_I are the covariance matrices for gradient and potential data respectively, and C_{GI} is their cross-covariance matrix.
U_G and U_I contain drift functions and F_G and F_I contain fault functions.
A, B, C, D are the solution of this linear system.
C^0_G is the covariance vector between the estimated increment and the gradient data and C^0_I is the covariance vector between the estimated increment and increment data.
U^0 and F^0 contain drift and fault functions at the estimated point.

Once the system has been solved, the iso-potential surface corresponding to the interface can be drawn. We can then visualize the 3D cube or cross-sections through it (Figure 1).

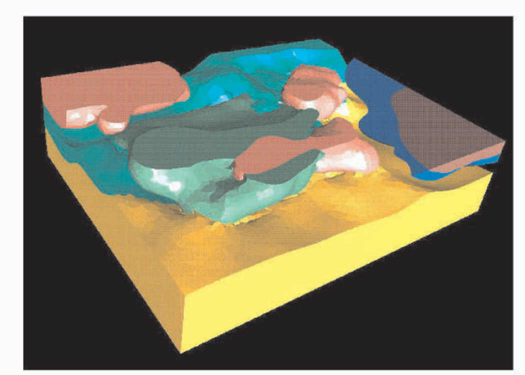


Figure 1. Example of a 3D geological model with the “Editeur géologique”

3 Variograms of orientation data illustrated by the Limousin dataset

The Limousin dataset, approximately a 70x70 km square, located in Centre France, is represented in Figure 2. Data sample a surface which is the top of a set of metamorphic rocks called “lower gneiss unit” (LGU). These data were all taken on the topographic surface.

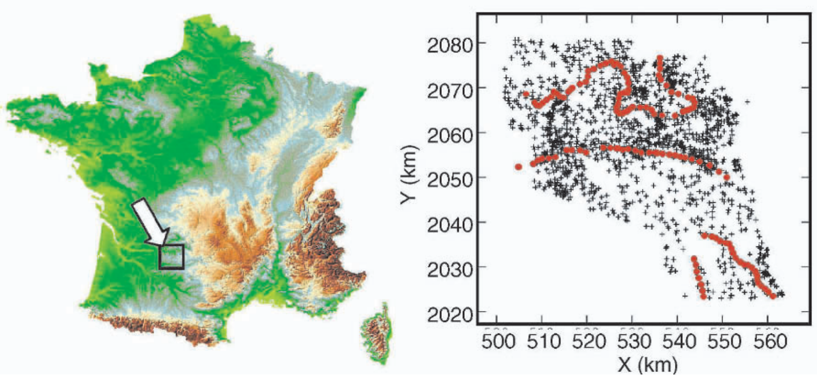


Figure 2. Base map of Limousin dataset. Black crosses: 1485 orientation data. Red discs: 133 interface data (digitized from the geological map).

3.1 ORIENTATION AND GRADIENT

Orientation data are vectors orthogonal to structural planes (e.g. foliation for metamorphic rocks, or stratification for sedimentary rocks) which are assumed to be parallel to the surfaces defined by the potential field. These data are sampled on the interface and also within the geological formations. They are considered as representing the gradient of the potential field. Since no intensity is usually attached to those gradient data, then vectors are arbitrarily considered as unit vectors. In practice, we work only with the three components of that vector. Let us mention that, in an orthonormal coordinate system, the mean of the components of a random unit vector is null and its variance is equal to one third.

3.2 COVARIANCE OF POTENTIAL AND GRADIENT

All the increments of potential are null, variograms of them are then useless. A first implementation of the method used a covariance given *a priori* by the user. But gradient data are algebraically linked with potential data. Therefore in this work, we use the only non-null data, namely gradient data, to infer the covariance model. Let $r = \sqrt{h_x^2 + h_y^2 + h_z^2}$ be the distance between two points and ${}'\mathbf{h} = (h_x, h_y, h_z)$ the vector joining these two points. Now, let K_P denote the covariance of Z and K_G^x, K_G^y, K_G^z the covariances of the three components of the gradient of Z. In order for Z to be differentiable, K_P must be twice differentiable (Chilès and Delfiner, 1999). Using the definition of differentiation, we can write the covariance of G^x, for instance:

$$K_{G^x}(\mathbf{h}) = -\frac{\partial^2 K_P(\mathbf{h})}{\partial h_x^{\;2}} \qquad \mathbf{(2)}$$

In the case of an isotropic covariance, $K_P(\mathbf{h}) = C(r)$ with C twice differentiable for $r \geq 0$ and we have:

$$K_G{}^x(\mathbf{h}) = -\left(\frac{C''(r)}{r^2} h_x{}^2 + C'(r)\left[\frac{1}{r} - \frac{h_x{}^2}{r^3} \right] \right) \qquad \mathbf{(3)}$$

More general formulas are available for anisotropic covariances.
The model parameters will be determined only with the sample variograms of the gradient data.
The cubic model with range *a* and sill *C₀*, chosen as basic model for K_P, is defined in the isotropic case by:

$$K_P(r) = C_0 \left(1 - 7\left(\frac{r}{a}\right)^2 + \frac{35}{4}\left(\frac{r}{a}\right)^3 - \frac{7}{2}\left(\frac{r}{a}\right)^5 + \frac{3}{4}\left(\frac{r}{a}\right)^7 \right) \quad \textit{for} \;\; 0 \leq r \leq a \qquad \mathbf{(4)}$$

$$K_P(r) = 0 \quad \textit{for} \;\; r \geq a$$

It is well adapted for geological contouring, because at the scale considered, geological surfaces are smooth and the cubic model has the necessary regularity at the origin.
Even if we assume the isotropy of K_P, K_G is necessarily anisotropic (Chauvet *et al.*, 1976). If we consider the partial derivative G^x for example, the extreme cases are the direction of the derivative, namely x, and the direction orthogonal to it, here y, and in term of variogram we have:

$$\gamma_{G^x}(h_y) = \frac{14C_0}{a^2}\left[\frac{15}{8}\frac{h_y}{a} - \frac{5}{4}\left(\frac{h_y}{a}\right)^3 + \frac{3}{8}\left(\frac{h_y}{a}\right)^5\right] \quad for \;\; h_y \le a \qquad \mathbf{(5)}$$

$$\gamma_{G^x}(h_x) = \frac{28C_0}{a^2}\left[\frac{15}{8}\frac{h_x}{a} - \frac{5}{4}\left(\frac{h_x}{a}\right)^3 + \frac{3}{8}\left(\frac{h_x}{a}\right)^5\right] - \frac{7C_0}{a^2}\left[5\left(\frac{h_x}{a}\right)^3 - 3\left(\frac{h_x}{a}\right)^5\right] \quad for \;\; h_x \le a \qquad \mathbf{(6)}$$

We recognize a pentaspheric model in the direction orthogonal to that of the partial derivative and a model with a hole effect in the direction of derivation.
In the other directions, the graph of the variogram is comprised between these two envelopes (Renard and Ruffo, 1993).

3.3 VARIOGRAM FITTING

For the Limousin case study, since the topography is rather smooth, the variograms have been computed in the horizontal plane only. Figure 3 shows sample variograms for the Limousin dataset.
The first remark is the difference of scale for the sill value between the vertical component and the horizontal ones. The reason is that the mean of the vertical gradient is significantly larger than zero due to the sub-horizontality of the layers, which results in a smaller variance for the vertical gradient component than for the horizontal ones. We also notice a large nugget effect for all components (nearly half of the total variability).
This difference of sill is modelled with a zonal anisotropy. The final model for the potential covariance is thus a nested cubic model:

$$K_P(\mathbf{h}) = K_3\left(\sqrt{h_x^2 + h_y^2 + h_z^2}\right) + K_2\left(\sqrt{h_x^2 + h_y^2}\right) + K_1\left(h_y\right) \qquad (7)$$

The ranges are 25000m, 17000m and 55000m, respectively, for K_3, K_2, and K_1.
The corresponding sills are 781000, 1700000, 10800000, respectively.
In comparison, the default values previously proposed by the software correspond to a single isotropic component with a range of 98000m (the size of the domain) and a sill of 229×10^6.
These two covariance models lead to rather different geometric models. For example, the depth of the LGU interface is up to 450m deeper with the default covariance model than with the fitted covariance. However, we must not forget, that we are extrapolating from data sampled on the topographic surface only.

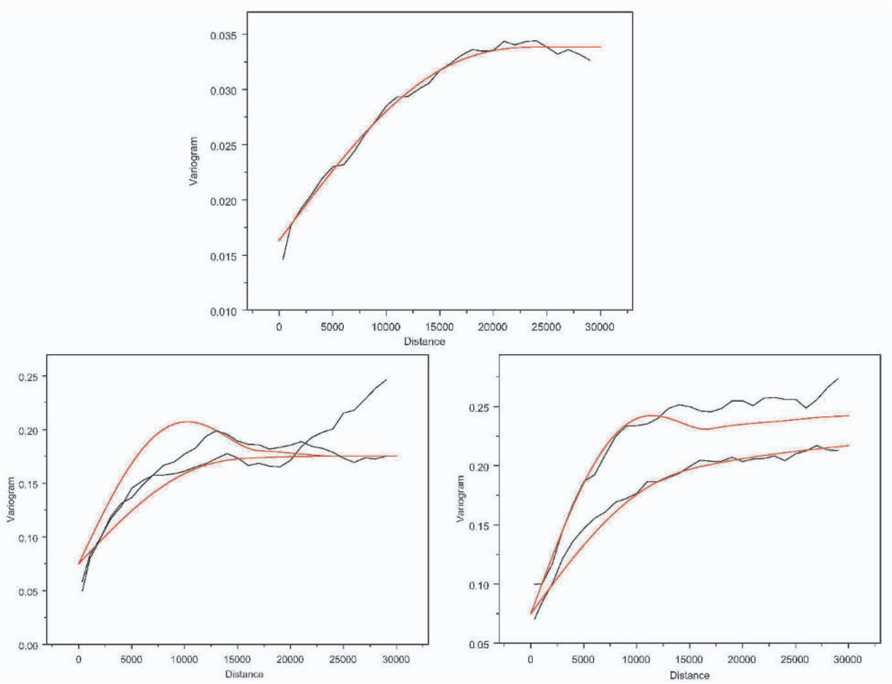


Figure 3. Experimental and fitted variograms for the components of the gradient. G^z (top), $G^{x//}$ and $G^{x\perp}$(bottom left) and $G^{y//}$ and $G^{y\perp}$ (bottom right). The symbol // (resp. ⊥) corresponds to the variogram in the direction of differentiation (resp. in a direction orthogonal to that of the differentiation).

In order to make the software easy to use for non-geostatisticians, an automatic procedure of variogram fitting based on the Levenberg-Marquardt method (Marquardt, 1963) has been implemented. The aim is the minimisation of the weighted metric distance between the sample variograms and the variogram model in the vectorial space of the fitted parameters (nugget effect, sill, range). It is a non linear regression method optimally using two minimisation approaches: quadratic and linear. A factor allows the use of one or another.

4 Determination of uncertainty

4.1 "REDUCED POTENTIAL" CARTOGRAPHY

When the covariance was chosen *a priori*, without consideration to a sample variogram, the method could not claim for optimality and the cokriging variance had no precise meaning. But now, since the model is well defined, determining the uncertainty on the position of the interface in depth makes sense and to get a better idea of the degree of uncertainty for the drawing we define a "reduced potential".
Let Z_0 be the value of the potential for a point on the considered interface, $Z^*(\mathbf{x})$ the value estimated at a point $\mathbf{x}$ and $\sigma_{CK}(\mathbf{x})$ the cokriging standard deviation at the same point.

The reduced potential $\Phi(\mathbf{x})$ is given by:

$$\Phi(\mathbf{x}) = \frac{Z^*(\mathbf{x}) - Z_0}{\sigma_{CK}(\mathbf{x})} \qquad (8)$$

For a given point, this variable represents the reduced estimation of the potential deviation from Z_0. It can be shown that the larger this value, the less chance the point has to be on the interface. With a Gaussian assumption for the potential field, Φ is a standardized normal variable, so that for example, the area inside the curves $\Phi = \pm 2$ includes the interface in about 95% of the cases. Figure 4 shows the interpolated LGU interface (black line) and the value of Φ in blue. In short, the yellow zone, which corresponds to $|\Phi| < 3$, is like a forbidden area for the drawing of the interface.

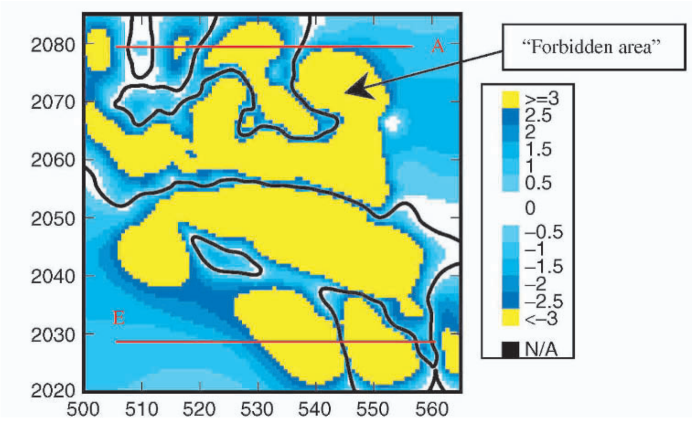


Figure 4. Limousin dataset. Map of the reduced potential.

Likewise, Figure 5 shows two cross-sections in the north (A) and the south (B) of the field. Of course, when the number of data is large, the position of the interface is well constrained, whereas in extrapolation there is a lot of uncertainty.

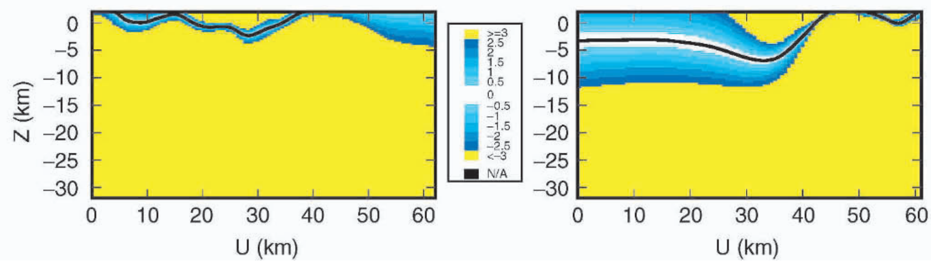


Figure 5. Limousin dataset. Cross-section A (left) and B (right) of the reduced potential.

4.2 UNCERTAINTY ON MODEL PARAMETERS

The covariance fitting has some part of uncertainty too. Thanks to a Bayesian approach it is possible to determine the uncertainty of the model parameters (Goria, 2004).
The aim is to simulate these parameters according to a posterior distribution, which is proportional to an *a priori* distribution and a likelihood function:

$$\pi(\theta \mid Z) \propto \pi(\theta)\pi(Z \mid \theta) \qquad (9)$$

where $\pi(\theta)$ is the *a priori* distribution of the parameters vector θ and $\pi(Z|\theta)$ is a likelihood function. The vector θ includes the coefficients of the drift basis function, and the parameters of the covariance.
We assume a normal distribution for the coefficient of the drift and a gamma distribution for the precision (inverse of the sill). For the range and the relative nugget effect, a discrete uniform prior is used. The results show a large uncertainty on the model parameters. If we use the maximum values of the estimated parameters for the covariance model, we see some differences in the geometry of the interface. For example, the depth of the LGU interface is around 200m deeper with this “Bayesian” covariance model than with the classical fitted covariance.
The posterior distribution can also be incorporated in the cokriging or conditional simulation process.

5 Other practical implementation issues

5.1 SEVERAL INTERFACES

When there are several geological layers, some rules must be respected to avoid crossing the boundaries. If the interfaces are not sub-parallel, several potential fields are used. Two rules, “erode” and “onlap”, as well as a stratigraphic column make it possible to solve all the issues facing us. The column defines the chronological order of the interfaces and the rules define the priority between the layers. The rule “erode” has always the priority and is used to mask the eroded part of the previous formations or to model an intrusive body. For example, on Figure 7 right, we can see the interface (1) which is in onlap on the interface (2).

5.2 FAULTS

Discontinuities are taken into account too. Faults are defined as external discontinuous drift functions in the cokriging system (Maréchal, 1984). The method requires the knowledge of the fault planes and the zones of effect of the faults. The discontinuity can be “infinite” and then crosses the whole field, dividing it in two subzones D and D’. The fault induces a discontinuity in the potential field, taken into account by a drift function such as:

$$f(\mathbf{x}) = 1_D(\mathbf{x})$$

This function complements the classical polynomial drift functions used for the non-stationarity in the cokriging system.
With a finite fault (Figure 6), the throw is determined with an influence area. As with an infinite fault, the discontinuity divides the delimited area in two sub-zones. In this case, the drift function has a bounded support and the function reaches its maximum at the centre of the fault. Outside the area the fault has no effect.

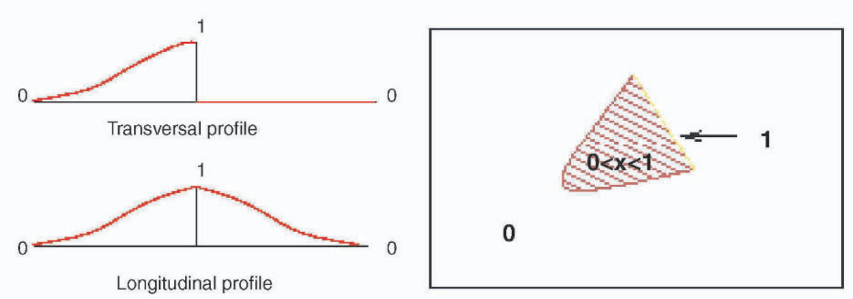


Figure 6. Finite fault. Left, transversal profile (top) and longitudinal profile (bottom) of the drift function. Right, area of influence of the discontinuity in the horizontal plane.

5.3 BOREHOLE ENDS

The last term in equation (1) is normally equal to zero, but could be strictly positive or negative if the points are not on the interface, which is the case when dealing with borehole ends. For example, the increment of potential is positive when the borehole end is above the considered interface with the following convention: the potential grows from the oldest geological formation to the most recent one.
Incomplete drillings can lead to a bad interpolation if a pre-processing of these soft data is not implemented. We use an iterative technique method based on the Gibbs sampler (Geman and Geman, 1984; Gilks *et al.*, 1996) to replace these soft data by hard data honouring both the inequalities and the spatial structure. That method, developed for stationary random functions (Freulon and de Fouquet, 1993) has been extended to the nonstationary case.
Figure 7 shows a synthetic example with two drill-holes (A and B) and two interfaces to be reconstructed (higher (1) with "onlap": 2 points and 1 gradient; lower (2) with rule "erode": 3 points and 1 gradient).

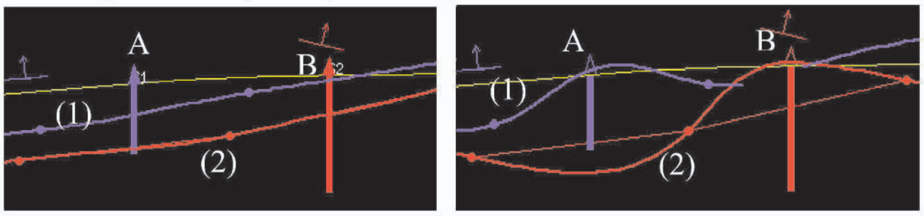


Figure 7. Interpolation of two interfaces. Left, without pre-processing, right with pre-processing.

The borehole A is only filled up with the facies whose interface (1) is the top and borehole B only with the facies whose interface (2) is the top.
On one hand, if borehole ends are not taken into account, the interpolation does not respect them as shown on Figure 7 (left).
On the other hand, Figure 7 (right) displays the result after pre-processing, interface (1) is in onlap as expected.
If P_1 and P_2 are respectively the iso-potential values for interface (1) and (2), and $P_b(A)$ and $P_e(A)$ respectively the values of potential at the beginning and at the end of the borehole, the result of the simulation gives values which respect $P_b(A)<P_1$ and $P_e(A)>P_2$.

6 Conclusion and future works

The potential-field method used in 3D geological modelling makes it possible to create models, even in complex situations, that combine different types of data, especially structural data. Thanks to the variography of these data it is possible to specify a sensible model of covariance and then to produce maps of uncertainty for the position of geological interfaces.
In this method we consider orientation data as gradient data, namely unit vectors. Only in specific cases, we know the structural intensity. The objective of ongoing work is to show, with simulations of actual situations, the impact on the covariance when actual gradient vectors are replaced by unit vectors.
Other improvements are planned like a better fault processing or the use of geophysical data.

References

Chauvet, P., Pailleux, J. and Chilès, J.-P., Analyse objective des champs météorologiques par cokrigeage, *La Météorologie, Sciences et Techniques, 6e série*, no. 4, 1976, p. 37-54.

Chilès, J.-P. and Delfiner, P., *Geostatistics: Modeling Spatial Uncertainty*, Wiley, 1999.

Freulon, X. and de Fouquet, C., Conditioning a Gaussian model with inequalities, *Geostatistics Troia '92*, vol. 1, A. Soares (Ed.), Kluwer, 1993, p. 201-212.

Geman, S. and Geman, D., Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, PAMI-6, no. 6, 1984, p. 721-741.

Gilks, W.R., Richardson, S. and Spiegelhalter, D.J., *Markov Chain Monte Carlo in practice*, Chapman and Hall/CRC, 1996.

Goria, S., *Evaluation d'un projet minier: approche bayésienne et options réelles*, Ph.D. Thesis, Ecole des Mines de Paris, 2004.

Lajaunie, C., Courrioux, G. and Manuel , L., Foliation fields and 3D cartography in geology : principles of a method based on potential interpolation, *Mathematical Geology*, vol. 29, no. 4, 1997, p. 571-584.

Maréchal, A., Kriging seismic data in presence of faults, *Geostatistics for Natural Resources Characterization*, Part 1, G. Verly *et al.* (Eds), Reidel, p. 271-294.

Marquardt, D., An algorithm for least-squares estimation of non-linear parameters. *SIAM, Journal on Applied Mathematics*, vol. 11, no. 2, 1963, p. 431-441.

Renard, D. and Ruffo, P., Depth, dip and gradient, *Geostatistics Troia '92*, vol. 1, A. Soares (Ed.), Kluwer, 1993, p. 167-178.

ACCOUNTING FOR NON-STATIONARITY AND INTERACTIONS IN OBJECT SIMULATION FOR RESERVOIR HETEROGENEITY CHARACTERIZATION

DENIS ALLARD[1], ROLAND FROIDEVAUX[2] and PIERRE BIVER[3]
1 *INRA, Unité de Biométrie, Site Agroparc, 84914 Avignon, France*
2 *FSS Consultants SA, 9, rue Boissonnas, 1227 Geneva, Switzerland*
3 *Total, 64018 Pau, France*

Abstract. This paper proposes an algorithm for simulating object models based on an underlying Markov point process able to reproduce attraction or repulsion between objects in a non stationary setting based on workable approximations for computing the local intensity, taking into account: i) non stationary proportions and non stationary object parameters; ii) erosion rules in the case of multi-type objects; iii) attraction or repulsion between objects.

1 Introduction

Modeling heterogeneity is the first, and possibly the most important, step of a reservoir characterization study. Depending on the geological context, several simulation techniques can be envisioned to perform this first step: sequential indicator simulation (Alabert, 1987; Goovaerts, 1997), transition probability simulation (Carle and Fogg, 1996), sequential simulation using multi-points statistics (Strebelle, 2002), truncated Gaussian or plurigaussian simulation(Le Loc'h and Galli, 1997), or Boolean simulation (Haldorsen and Macdonald, 1987; Lantuéjoul, 2002). An important feature of object simulation, which sets it apart from the other techniques, is the fact that it is not pixel-based, i.e it does not generate values at the nodes of a pre-defined grid. Rather, it generates geometric shapes in space according to some probability laws.

Although Boolean model simulation has been widely used during the last two decades to simulate sedimentary bodies (especially in fluvio-deltaic environments), several non trivial issues have remained and require scrutiny. Any general purpose object simulation program for reservoir characterization should (i) allow for the simulation of multiple object types, (ii) respect user-defined erosion rules between object of different types, (iii) reproduce specified a priori proportions, after erosion, for each object type, (iv) account for non-stationary object dimensions and orientations, (v) be conditional to existing hard-data, (vi) account for inter-actions (attraction or repulsion) between objects.

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 155-164.

A key issue is that the single most important parameter for object models simulation, namely the (non stationary) intensity of the underlying object process, is not a parameter provided by the end user but must instead be inferred from the other input parameters listed above. A second very important issue is the fact that Boolean models traditionally used assume independence between objects. As such they are inadequate for reproducing interactions between objects. Lantuéjoul (1997, 2002) proposed a birth and death process for simulating conditional Boolean models with non stationary intensity. This algorithm considers the intensity as known and does not address the problem of making a bridge between intensity and local proportion. Recently, Benito García Morales (2003) proposed a method based on Wiener filter to estimate a non stationary intensity from non stationary proportions. This method assumes stationary distribution function of the object parameters. None of the methods described above consider multi-type objects.

This paper proposes an algorithm for simulating multi-type object models based on an underlying Markov point process able to reproduce attraction or repulsion between objects in a non stationary setting.

2 Boolean Models

2.1 GENERAL OVERVIEW

A single object type Boolean model (Stoyan, Kendall and Mecke, 1995) is made of two parts:

- A set of points (seeds), denoted $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$, which follow a Poisson distribution characterized by its intensity θ describing the expected number of object centroïds per unit volume. This intensity may be varying in space. As a consequence of the Poisson assumption, object centroïds are independent to each other.
- Random variables, independent of the Poisson process, that attach to each of these points random marks describing the shape, dimensions and orientation of objects A. These random variables are described by their joint probability density ψ. The random marks are independent one from the other.

The key parameters for Boolean models simulation is the point intensity parameter θ which describes how many object centroïds are expected per volume unit. This parameter, however, is not readily available: geologists have good ideas about the proportion for each geological object they want to simulate, but they have no feel for the number of such objects. Hence the need to estimate the intensity θ from the proportion p. In stationary conditions, it is well known that for Boolean models the proportion is related to the intensity according to the following relationship (Lantuéjoul, 2002; Stoyan *et al.*, 1995):

$$p = 1 - \exp\left\{-\theta \int_{\mathbf{R}^3} E_\psi[\mathbf{1}_{0 \in A(\mathbf{v})}]\, d\mathbf{v}\right\} = 1 - \exp\{-\theta V\},$$

where $\mathbf{1}$ is the indicator function, 0 is the origin, $A(\mathbf{v})$ is a random object centered in $\mathbf{v}$ and V is the expectation of the volume of a random object A whose mark density is ψ. Inverting this relationship yields to

$$\theta = -\frac{1}{V}\ln(1-p). \tag{1}$$

In practice, this congenial situation is the exception rather than the rule: a priori proportions and mark densities are not stationary, objects of different types do not overlap each other randomly but according to erosion rules, objects of a given type may show a tendency to attract each other or, conversely, to repulse each other. In all these situation equation (1) cannot be used directly but requires adjustments.

2.2 ACCOUNTING FOR EROSION RULES

For each object type $k = 1, \ldots, K$, there is a corresponding proportion p_k, intensity θ_k and mark density ψ_k. Although equation (1) already accounts for the fact that several objects of the same type may overlap, it requires adjustment to ensure that, in case of multiple object type simulation, the target proportion of each type is correctly reproduced. In practice, this is done by substituting in equation (1) the proportion p_k by a corrected proportion p'_k. This correction depends on the "erosion rule" determining which type of object erodes the other. Among all the possible rules, three are commonly used: *random overlapping*, the *vertical erosion* rule (the object with the highest centroïd erodes the others) and the *hierarchical erosion* rule whereby the object type 1 always erodes the object type 2 which, in turn, always erodes object type 3, etc.

- In the case of an hierarchical erosion rule, the proportion of the type 1 object does not need to be corrected. Type 2 objects will be partly eroded by objects of type 1. Hence, for a visible proportion p_2, a corrected proportion $p'_2 = p_2/(1-p_1)$ needs to be simulated. Recursively, for the type k, the corrected proportion is given by:
$$p'_k = \frac{p_k}{1-\sum_{i=1}^{k-1} p_i}. \tag{2}$$
- Derivation of a corrected proportion for a vertical erosion rule is somewhat more complicated. A second order approximation of this corrected proportion is given by:
$$p'_k = p_k\left(1+\frac{(1+p_{tot})(p_{tot}-p_k)}{2}\right), \tag{3}$$
where $p_{tot} = \sum_k p_k$ is the total proportion of objects. This approximation relies on the idea that in the case of a vertical erosion, it is equally likely that an object of type k erodes an object of type l than the opposite. It leads to a corrected proportion $p'_k < 1$.
- In the case of random overlapping between object types, it is also equally likely that an object of type k overlaps an object of type l than the opposite. Hence, the same corrections as those used for vertical erosion are used.

2.3 ACCOUNTING FOR NON STATIONARITY

If object proportions or object parameters are non stationary the relationship between the non stationary proportions and non stationary parameters is more complex. Dropping, for ease of notation, the subscript referring to the object type, this relationship is expressed locally at a point $\mathbf{u} \notin \mathbf{X}$ as:

$$p(\mathbf{u}) = 1 - \exp\left\{ - \int_{\mathbf{R}^3} \theta(\mathbf{v}) E_{\psi(\mathbf{v})}[\mathbf{1}_{\mathbf{u} \in A(\mathbf{v})}]\, d\mathbf{v} \right\}, \tag{4}$$

where the expectation is computed with respect to the mark density with local parameters $\psi(\mathbf{v})$. This expression is extremely difficult (if not impossible) to invert. If $\psi(\mathbf{u})$ and $\theta(\mathbf{u})$ are smooth and slowly varying functions, then first order expansion in equation (4) can be used locally to approximate the local intensity $\theta_k(\mathbf{u})$ from the local corrected proportion $p'_k(\mathbf{u})$:

$$\theta_k(\mathbf{u}) = -\frac{1}{V_k(\mathbf{u})} \ln(1 - p'_k(\mathbf{u})), \tag{5}$$

where $V_k(\mathbf{u})$ is the local expectation computed using the local mark density $\psi_k(\mathbf{u})$ and $p'_k(\mathbf{u})$ is the proportion corrected to account for erosion as described above.

3 Markov object models

It is sometimes necessary to impose that objects of a given family are attracted to each other or on the contrary that there is some sort of repulsion between objects. The general idea is to consider that repulsion or attraction is a feature of the underlying point processes, but that marks are still independent from each other. The appropriate framework for such point processes is the Markov point processes (MPP). Poisson point processes on which are built Boolean models is a particular case of MPP, for which there is no repulsion and no attraction. A comprehensive presentation of MPP can be found in Stoyan *et al.* (1995) or van Lieshout (2000).

3.1 GENERAL PRESENTATION OF MARKOV POINT PROCESSES

Markov point processes are point processes for which points are no longer independent from each other but are dependent on the configuration of the other points. According to the Hammersley-Clifford theorem, the probability density function (pdf) of a MPP depends only on functions of *cliques*. Cliques are set of points such that each point of this set is a neighbor of all other points of the set. The neighborhood relationship, used to define cliques must be symmetrical. Usually, the points $\mathbf{x}$ and $\mathbf{y}$ are neighbors (denoted $x \sim y$) if their distance $d(\mathbf{x}, \mathbf{y})$ is less than R for some $R > 0$.

The simplest possible clique to consider is a clique consisting of a single point. In this case there is no interaction and we are back to the classical Poisson process framework. In order to account for interaction, cliques of more than one point need to be considered. In practice, two point cliques will be considered and pairwise interaction functions, denoted $\beta(\mathbf{x},\mathbf{y})$, will be used to define the density of a configuration $\mathbf{X}$:

$$f(\mathbf{X}) \propto \prod_{\mathbf{x}\in\mathbf{X}} \theta(\mathbf{x}) \prod_{\mathbf{x},\mathbf{y}\in\mathbf{X}\,:\,\mathbf{x}\sim\mathbf{y}} \beta(\mathbf{x},\mathbf{y}),$$

where $\theta(\mathbf{x})$ is the intensity function. Among all pairwise interaction point processes, the simplest one is the Strauss process (Strauss, 1975; Kelly and Ripley, 1976) for which the interaction function is constant:

$$\beta(\mathbf{x},\mathbf{y}) = \beta \text{ if } \mathbf{x}\sim\mathbf{y} \text{ and } \beta(\mathbf{x},\mathbf{y}) = 1 \text{ otherwise,} \tag{6}$$

with $0 \le \beta \le 1$. The pdf of a Strauss process is thus

$$f(\mathbf{X}) = \alpha\beta^{n(\mathbf{X})} \prod_{\mathbf{x}\in\mathbf{X}} \theta(\mathbf{x}),$$

where α is the normalizing constant and $n(\mathbf{X})$ is the number of neighbor pairs $\mathbf{x},\mathbf{y}\in\mathbf{X}$ with respect to the neighborhood relationship.

- If $\beta = 1$, there is no interaction whatsoever, and we are back to the non stationary Poisson point process with intensity $\theta(\mathbf{u})$.
- If $0 \le \beta < 1$, there is some repulsion. Configurations with a high number of neighbors have a smaller density than configurations with a low number of neighbors. As a result the point process is more regular than a Poisson point process. In particular, if $\beta = 0$, configurations with neighbors have a null density and are thus impossible.
- The case of an attraction would correspond to $\beta > 1$, but without additional constraints it is mathematically not admissible because the associated density does not integrate to a finite quantity (Kelly and Ripley, 1976). However, the interaction function

  $$\beta(\mathbf{x},\mathbf{y}) = \beta \text{ if } r \le d \le R,\ \beta(\mathbf{x},\mathbf{y}) = 0 \text{ if } d < r \text{ and } \beta(\mathbf{x},\mathbf{y}) = 1 \text{ if } d > R, \tag{7}$$

  where $0 < r < R$ and d stands for $d(\mathbf{x},\mathbf{y})$, is an admissible model. In practice, the restriction introduced by r is not important because r can be chosen arbitrarily small. A typical choice is the mesh of the grid on which the simulation is represented.

In the following we will consider Strauss models for both repulsion and attraction, with the additional condition on r for attraction. The conditional density of adding to the configuration $\mathbf{X}$ a point in $\mathbf{u}$ is

$$f(\mathbf{X}\cup\{\mathbf{u}\} \mid \mathbf{X}) = f(\mathbf{X}\cup\{\mathbf{u}\})/f(\mathbf{X}) = \theta(\mathbf{u})\beta^{n(\partial\mathbf{u})}, \tag{8}$$

where $n(\partial\mathbf{u})$ is the number of neighbors of $\mathbf{u}$. Hence, the parameter β can be interpreted as a factor multiplying locally the intensity for each point in the neighborhood of $\mathbf{u}$.

3.2 DERIVING THE INTENSITY FOR OBJECT MODELS BASED ON STRAUSS PROCESSES

Objects are now attached to the Markov point process, and for sake of simplicity, we first consider the stationary case. The proportion of objects attached to Markov point processes is not easily related to the intensity: there is no relationship comparable to (4). In the case of a Strauss model, a workable approximation for the intensity was found to be:

$$\theta(\mathbf{u}) = -\frac{p'(\mathbf{u})}{V(\mathbf{u})}\left(1 + c\frac{p'(\mathbf{u})}{2}\right), \tag{9}$$

where $p'(\mathbf{u})$ is the proportion corrected for the erosion (as described in Section 2.2) and c is approximately the conditional probability that $\mathbf{u}$ is in an object A' given that it is already in an object A. In reservoir simulations, the objects have generally random size to account for the natural variability of geological objects. For the direction i, let us denote X_i the dimensions of an object, R_i the dimension of the interaction box and r_i the minimal distance in case of attraction. Conditional on R_i, $i = 1, 2, 3$, it can be shown that

$$c = \beta(1 - \frac{r_1 r_2 r_3}{X_1 X_2 X_3}) + (1-\beta)(1 - \frac{R_1 R_2 R_3}{X_1 X_2 X_3})[1 - \mathbf{1}_B(X_1, X_2, X_3)], \tag{10}$$

where $\mathbf{1}_B(X_1, X_2, X_3)$ is the indicator function of the vector (X_1, X_2, X_3) being in the box B defined by the dimensions (R_1, R_2, R_3).

Markov point processes are usually defined for fixed interaction distances, as in Section 3.1 and objects are usually random with probability functions F_i. Taking the expectations of Equation (10) leads to

$$c = \beta g(r_1, r_2, r_3) + (1-\beta) g(R_1, R_2, R_3), \tag{11}$$

with

$$\begin{aligned} g(R_1, R_2, R_3) &= 1 - F_1(R_1)F_2(R_2)F_3(R_3) - R_1 R_2 R_3 h_1(a_1) h_2(a_2) h_3(a_3) \\ &\quad - [h_1(a_1) - h_1(R_1)]\,[h_2(a_2) - h_2(R_2)]\,[h_3(a_3) - h_3(R_3)]), \end{aligned} \tag{12}$$

where a_i is the smallest dimension of the object in the direction i and $h_i(r_i) = \int_{\max\{a_i, r_i\}}^{\infty} f_i(x)/x\,dx$.

4 Simulation using birth and death processes

Non conditional Boolean models (i.e. corresponding to $\beta = 1$) can be simulated directly: for each type of object k, first draw the number of objects from a Poisson random variable with parameter $\Theta_k = \int_D \theta_k(\mathbf{u})\,d\mathbf{u}$, then locate randomly the objects according to the intensity $\theta_k(\mathbf{u})$.

In all other cases (presence of conditioning data and/or Markov object models) simulation must be performed using a birth and death process. Birth and death processes are continuous time Markov Chains belonging to the family of Markov Chain Monte Carlo (MCMC) methods (van Lieshout, 2002; Lantuéjoul, 2002).

Starting from an initial configuration, an object is either removed or added according to some transition probability that depends on the current configuration of the simulation at each time step. This transition probability is chosen in such a way that the stationary distribution of the Markov chain is precisely the density we wish to simulate from. According to standard results of Markov chain theory, if the birth and death process is ergodic, there exists a stationary spatial distribution and the convergence to the stationary distribution will always occur independently on the initial configuration. Ergodicity holds if the detailed balance equation is verified at each iteration (see e.g. van Lieshout, 2002 p. 79).

It can be shown that the detailed balance is verified for the following choices: the probability of choosing a birth is $q(\mathbf{X}) = \Theta/(\Theta + n(\mathbf{X}))$ where Θ is the sum of $\theta(\mathbf{u})$ on $\mathcal{D}$. Then, $1 - q(\mathbf{X})$ is the probability of choosing a death. For a Strauss point process with $\beta \neq 1$, it is convenient to introduce an auxiliary field $\sigma(\mathbf{u})$ defined in the following way: for a repulsion (i.e., $\beta < 1$), $\sigma(\mathbf{u}) = \beta^{n(\partial\mathbf{u})}$; for an attraction (i.e., $\beta > 1$), $\sigma(\mathbf{u}) = \beta^{\min\{(n(\partial\mathbf{u}) - n_{max}),0\}}$, where n_{max} is the maximum number of neighbors of each object. Its main effect is to stabilize the algorithm by avoiding a large quantity of objects piling on each other without increasing the proportion of this object type. In case of birth, a new object is proposed in $\mathbf{u}$ proportionally to an intensity $b(\mathbf{X}, \mathbf{u}) = \theta(\mathbf{u})\sigma(\mathbf{u})$. In case of death, the object to be removed is chosen with a uniform probability among the list of objects.

For performing conditional multi type conditional simulations, the conditioning taking into account the erosion rules must be checked each time a new object is added or removed.

5 Implementation

The implementation of the algorithm described in the previous section raises some critical issues.

- *Border effects :* It is important to ensure that objects intersecting the domain D but whose centroïds are outside this domain can be simulated. A practical way consists in considering a bigger domain D^s whose dimension is the dimension of the domain under study, D, increased by the dimension of the largest conceivable object. There is one such domain D^s_k for each type of object and the expected number of objects of type k to be simulated must be computed on D^s_k. Intensities yust be extrapolated on the domain D^s_k not in D.
  For models with interaction, care must be taken to simulate correctly the Markov point process near the borders. By construction there cannot be any neighbors outside D^s. For a point $\mathbf{u}$ located near the border the number of neighbors $n(\partial\mathbf{u})$ will therefore be underestimated as compared to points located in the center of D^s. As a consequence, the field $\sigma(\mathbf{u})$ accounting for the interaction will be biased towards less interaction near the borders. In the case of repulsion for example, this bias results in an accumulation of objects near the border of D^s. Because the border of the augmented domain D^s is usually

outside the true domain D, the ultimate bias is less object than expected in D and a proportion under the target. To account for this bias, the number of neighbors is corrected by $n(\partial\mathbf{u})^* = n(\partial\mathbf{u})/v(\mathbf{u})$ where $v(\mathbf{u})$ is the proportion of the volume of the interaction box contained in D. Note that in case of mutiple object types, there is one such correction per object type.

- *Initial configuration :* In case of conditional simulation, an initial simulation is performed which will honor all the hard data. This is achieved by defining a new domain D^i which guarantees that any new simulated object, whatever its location, dimensions or orientation, will intersect at least one conditioning data. There is no birth and death process in this initial phase: new objects are added until all hard data are intersected. To avoid possible endless iterations, a maximum number of iterations is specified for this phase.
- *Convergence :* The question of finding a criterion for deciding if the algorithm has reached convergence is a very difficult one. There is no general rule for evaluating the number of iterations necessary to reach a pre-specified distance between the theoretical stationary distribution, and the actual distribution after n iterations. A considerable amount of literature has been devoted to this subject, see e.g. Meyn and Tweedie (1993) for a survey on this subject. Most of the proposed methods are either limited to some very simple cases or difficult and time consuming to implement. As a result, for practical purpose, the stopping rule will be a combination of a maximum number of iterations and a monitoring of some important output parameters (number of simulated objects, number of conditioning objects that have been replaced, average number of neighbors).

6 Illustrative example

To illustrate the proposed algorithm, let us consider tow examples. In both cases two types of objects are considered: dunes (fan shaped sedimentary bodies) and sinusoidal channels. In the first example, vertical proportion curves are imposed. For the dunes the proportion decreases steadily from a maximum of 30% at the top of the reservoir to a minimum value of 1% at the bottom. For the sinusoidal channels the trend is reversed: 30% at the bottom and 1% at the top. There are no interaction between the objects, neither for the channels nor for the dunes, and a vertical erosion rule is enforced. Figure 1 shows a typical cross-section of one realization. As can be seen, the trends in proportions are correctly reproduced. On average, the simulated proportion is 16% for the dunes and 14.9% for the channels, almost identical to the target proportion which were 15.5% in both cases. In the second example the target proportions are stationary, 10% for both the channels and the dunes, but object interactions are imposed: the dunes will repulse each other (the interaction box is 10% percent larger than the size of the objects) whereas the channels will attract each other. The interaction box is twice as large as the object width and height but has the same length, which means that the attraction operate only laterally and vertically. In this case a hierarchical erosion rule is applied. Figure 2 shows a horizontal and vertical cross-section. The simuated proportions are 10.5% for the fans and 10.0% for the channels. The results conform to the constraints which have been imposed: dunes are all distinct one from the other with no overlap and they systematically erode the channels, which tend to cluster together. Again, the average simulated proportions match almost exactly the target proportions.

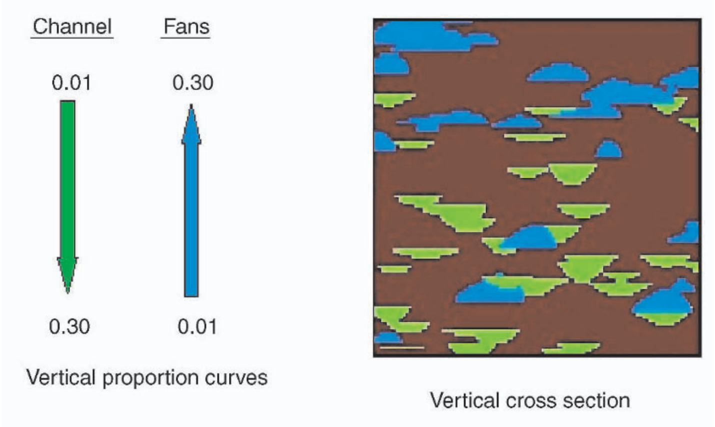


Figure 1. Vertical cross section of a simulation with vertical erosion rule

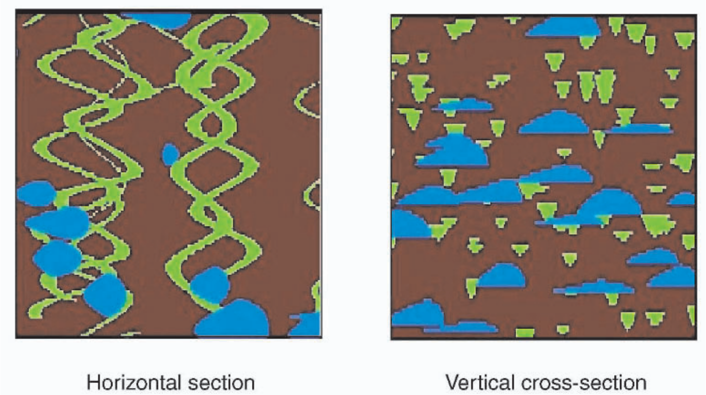


Figure 2. Horizontal and Vertical cross section of a simulation with hierarchical erosion rule

7 Discussion

The algorithm which has been presented offers a lot of flexibility and has proven effective for producing realistic simulations of reservoir heterogeneity in non-stationary situation. However, as is the case with all simulation algorithms, it is not universal and its limits of application should be respected. A critical issue, when performing conditional simulation, is the consistency between hard data, object parameters (shape and dimensions) and target proportions. This consistency becomes even more important if some or all parameters are non-stationary. The size of the object to be simulated is a critical issue. The larger the object the more difficult it will be to reproduce a target proportion and to honour conditioning data. Consistency between the object size (in particular its thickness) and the resolution at which facies are coded in well data must imperatively be verified. The approximations presented above are valid for not too high proportions. It is recommended that each proportion does note exceed 50% and that there is at least 20% of matrix, even locally. Although this algorithm can accommodate non-stationarity care should be taken that this non stationarity describes a smooth variation. Discontinuities must be avoided. The concept of neighborhood is not an intuitive one. Selecting too large a neighborhood may prove self-defeating: every point is the neighbour of every other point (they all belong to the same clique!). In case of attraction the points will not show any tendency to group in cluster, and in case of repulsion the process may never converge since it will be impossible to reach the target proportion and remain consistant with the repulsion constraint.

References

Alabert, F., 1987, Stochastic Imaging of Spatial Distribution Using Hard and Soft Information. M. Sc. Thesis, Stanford University.

Benito García-Morales, M., 2003, Non stationnarité dans les modèles de type Boléen: application à la simulation d'unités sédimentaires. PhD Thesis from the Ecole Nationale Supérieure des Mines de Paris, Centre de Géostatistique.

Carle, S.F. and Fogg G.E ., 1996, Transition probability-based indicator geostatistics: Math. Geology, v. 28, no. 4, p. 453–476.

Goovaerts, P., 1997, Geostatistics for natural resources evaluation: Oxford University Press, Oxford, 483 p.

Haldorsen, H.H. and Macdonald, C.J., 1987, Stochastic modeling of underground reservoir facies (SMURF). SPE 16751, p. 575–589.

Kelly, F. and Ripley, B.D., 1976, A note on Strauss's model for clustering: Biometrika, v. 63, no. 2, p. 357–360.

Lantuéjoul, C., 1997, Iterative algorithms for conditional simulations, *in* Baafi E.Y. and Schofield, N.A., eds, Geostatistics Wollongong '96: Kluwer Academic Publishers, Dordrecht, p. 27–40.

Lantuéjoul, C., 2002, Geostatistical simulation; models and algorithms: Springer-Verlag, Berlin, 256 p.

Le Loc'h, G. and Galli, A., 1997, Truncated plurigaussian method: theoretical and practical points of view, *in* Baafi E.Y. and Schofield, N.A., eds, Geostatistics Wollongong '96: Kluwer Academic Publishers, Dordrecht, p. 211–222.

Meyn, S.P. and Tweedie, R.L., 1993, Markov Chains and Stochastic Stability: Springer-Verlag, London, 550 p.

Stoyan, D., Kendall, W. and Mecke J., 1995, Stochastic Geometry and its Applications, 2nd Edition: John Wiley, Chichester, 436 p.

Strauss, D.J., 1975, A model for clustering: Biometrika, v. 62, no. 2, p. 467–475.

Strebelle, S., 2002, Conditional simulation of complex geological structures using multi-points statistics: Math. Geology, v. 34, no. 1, p. 1–22.

van Lieshout, M.N.N., 2002, Markov Point Processes: Imperial College Press, London, 175 p.

ESTIMATING THE TRACE LENGTH DISTRIBUTION OF FRACTURES FROM LINE SAMPLING DATA

CHRISTIAN LANTUÉJOUL, HÉLÈNE BEUCHER, JEAN-PAUL CHILÈS,
CHRISTIAN LAJAUNIE and HANS WACKERNAGEL
Ecole des Mines, 35 rue Saint-Honoré, 77305 Fontainebleau, France

PASCAL ELION
ANDRA, 1-7 rue Jean Monnet, 92298 Châtenay-Malabry, France

Abstract. This paper deals with the estimation of the length distribution of the set of traces induced by a fracture network along an outcrop. Because of field constraints (accessibility, visibility, censorship, etc...), all traces cannot be measured the same way. A measurement protocol is therefore introduced to systematize the sampling campaign. Of course, the estimation procedure must be based on this protocol so as to prevent any bias. Four parametric procedures are considered. Three of them (maximum likelihood, stochastic estimation-maximization and Bayesian estimation) are discussed and their performances are compared on 160 simulated data sets. They are finally applied to an actual data set of subvertical joints in limestone formations.

1 Introduction

Fractures such as faults and joints play a key role in the containment of nuclear waste in geological formations, in the oil recovery of a number of petroleum reservoirs, in the heat recovery of hot dry rock geothermal reservoirs, in the stability of rock excavations, etc. The fracture network is usually observable through its intersection with boreholes or through its traces on outcrops (see Fig. 1). An important parameter of a fracture network is the fracturation intensity, i.e. the mean area occupied by the fractures per unit volume, which is experimentally accessible even from unidimensional observations such as boreholes. The same fracturation intensity can however correspond to very different situations, whose extremes are a network of few large well-connected fractures and a network with a large number of small disconnected fractures. Getting geometrical and topological information about the fractures - size, orientation, aperture, connectivity - is therefore very important. Despite substantial work, this remains an arduous task, mainly because of the geometrical or stereological biases resulting from this limited observability (Chilès and de Marsily, 1993).

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 165-174.

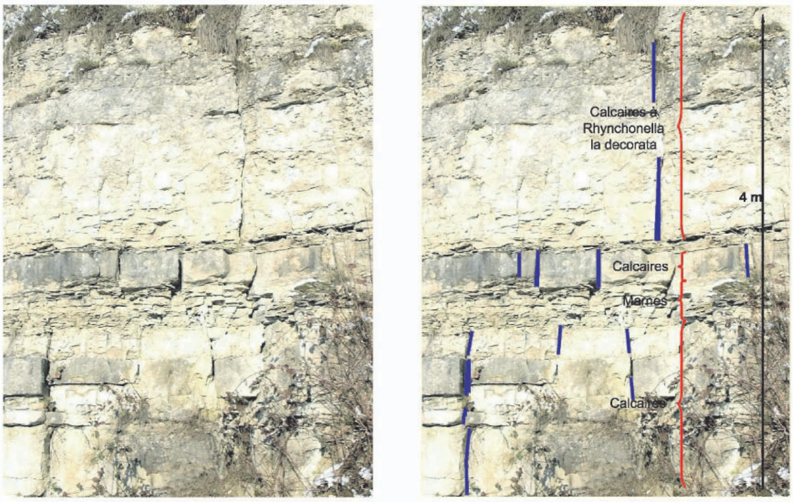


Figure 1. Example of an outcrop and its traces (vertical outcrop in a quarry, Oxfordian limestones, East of France)

This paper deals with the estimation of the length distribution of a set of traces along an outcrop. (The link between trace length and fracture size is briefly discussed at the end of the paper.) The difficulty lies in that the traces are usually not entirely visible. Their lower part is often buried. Their upper part may not be available either if the region around the outcrop has been eroded or mined out. In the practical case considered, the outcrop is a vertical face in a limestone quarry and all traces are sub-vertical.

In order to reduce risks associated with the sampling of such traces, it is convenient to resort to a sampling protocol that says what traces should be effectively measured and how. In the practical case considered, all traces are sub-vertical, which simplifies the protocol as well as its presentation (see Fig. 2):

1. All traces hitting a horizontal reference line - and only those traces - are selected for measurement;
2. Only the part of a selected trace above the reference line is actually measured;
3. A measurement is achieved even if the upper part of the trace is incomplete.

In other words, not all traces are sampled. Moreover, sampling a trace consists of measuring its residual - and sometimes censored - length above the reference line. The approach presented here is applicable to more general situations (the assumption that the traces are vertical or parallel is not really required), and can be easily generalized to other sampling schemes, including areal sampling.

In this paper, four parametric procedures[1] are proposed to estimate the trace length distribution starting from residual length data. These are the maximum likelihood estimation (MLE), its estimation-maximization variation (EM), the

[1] A non-parametric procedure based on the Kaplan-Meier estimation can also be designed. It is not described here to simplify the presentation.

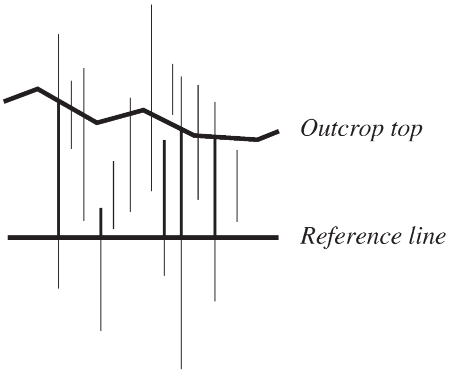


Figure 2. Protocol for sampling the outcrop

stochastic estimation-maximization algorithm (SEM) and the Bayesian estimation (BE). The performances of three of them (MLE, SEM and BE) are compared on data sets simulated from a Weibull distribution. Finally, the same three procedures are applied to a fracturation data set coming from an underground research laboratory of ANDRA (French national radioactive waste management agency).

2 Consequences derived from the protocol

At first, it should be pointed out that the selection of the traces is biased. The longer a trace, the more chance it has to hit the reference line. Quantitatively speaking, if $g(\ell)$ denotes the probability density function (p.d.f.) of the traces with length ℓ, then that of the selected traces is $\ell g(\ell)/m$. In this formula, the mean m of g acts as a normation factor.

Now a selected trace hits the reference line at an arbitrary location. In probabilistic terms, this amounts to saying that a residual length can be written as a product LU, where L is the length of a selected trace and U is a random variable uniformly distributed on $]0,1[$ and independent of L. Accordingly, the cumulative distribution function (c.d.f.) R of the residual lengths satisfies

$$1 - R(\ell) = P\{LU \geq \ell\} = \int_{\ell}^{+\infty} \frac{xg(x)}{m} P\Big\{U \geq \frac{\ell}{x}\Big\} dx = \frac{1}{m} \int_{\ell}^{+\infty} (x - \ell)\, g(x)\, dx.$$

By a first differentiation, the p.d.f. r of the residual lengths is obtained as a function of the c.d.f. G of the actual traces

$$r(\ell) = \frac{1 - G(\ell)}{m}, \tag{1}$$

and by a second differentiation both p.d.f.'s turn out to be related by the formula

$$g(\ell) = -\frac{r'(\ell)}{r(0)}\cdot \tag{2}$$

3 Four estimation procedures

We turn now to the problem of estimating the p.d.f. g of the actual traces starting from the available residual traces, namely the complete ones $\ell_I = \big(\ell_i, i \in I\big)$ and the censored ones $t_J = \big(t_j, j \in J\big)$. Equation (2) suggests to concentrate at first on the estimation of r, and then on that of g. The four procedures presented hereunder have been designed along that line.

3.1 MAXIMUM LIKELIHOOD ESTIMATION (MLE)

In this procedure, the trace length p.d.f. g is supposed to belong to a parametrized family $\big(g(\cdot|\theta), \theta \in T\big)$. For each p.d.f. $g(\cdot|\theta)\}$, a residual p.d.f. $r(\cdot|\theta)$ can be associated. The MLE procedure consists of finding a parameter θ that maximizes the likelihood of the data

$$L(\ell_I, t_J, \theta) = \prod_{i \in I} r(\ell_i|\theta) \prod_{j \in J} [1 - R(t_j|\theta)]$$

or

$$L(\ell_I, t_J, \theta) = r(\ell_I|\theta)[1 - R(t_J|\theta)] \tag{3}$$

for short (Laslett, 1982). It should be pointed out that this procedure is not universal. For instance, the likelihood may have no maximum[2]. Moreover, even if a maximum does exist, its determination by differentiation of the likekihood may turn out to be ineffective.

3.2 EXPECTATION-MAXIMIZATION PROCEDURE (EM)

A possible approach for estimating the maximum likelihood is to resort to the EM algorithm (Dempster *et al.*, 1977). This is an iterative algorithm that produces a sequence of parameter values in such a way that the likelihood of the data increases at each iteration. To present this algorithm, it is convenient to introduce the residual random lengths $L_J = \big(L_j, j \in J\big)$ that have been censored to $t_J = \big(t_j, j \in J\big)$. Of course $L_J \geq t_J$.

```
(i) let $\theta$ be the current parameter value;
(ii) calculate the conditional distribution $r_\theta$ of $L_J$ given $L_J \geq t_J$;
(iii) find $\theta_m$ that maximizes $\theta' \longrightarrow E_{r_\theta} \ln[r(\ell_I|\theta') r(L_J|\theta')]$;
(iv) put $\theta = \theta_m$, and goto (ii).
```

It should be pointed out that step (iii) of this algorithm also includes a maximization procedure. However the functions to be maximized do not depend on R and are therefore simpler to maximize than the likelihood of the censored data.

Nonetheless, this algorithm has some drawbacks. Calculating the expectation of step (iii) may be problematic. On the other hand, convergence may take place only to a local maximum that depends on the initial parameter value. Finally, the rate of convergence may be quite slow.

[2] However, the family of p.d.f.'s is usually designed so as to warrant a maximum whatever the data set considered.

3.3 STOCHASTIC EXPECTATION-MAXIMIZATION PROCEDURE (EM)

All these difficulties prompted Celeux and Diebolt (1985) to introduce the SEM algorithm that consists of replacing the calculation of the expectation by a simulation:
(i) let θ be the current parameter;
(ii) generate $\ell_J \sim r_\theta$;
(iii) find θ_m that maximizes $\theta' \longrightarrow r(\ell_I|\theta')r(\ell_J|\theta')$;
(iv) put $\theta = \theta_m$, and goto (ii).
Once again, this algorithm requires a maximization procedure. But what has to be maximized is the likelihood of pseudo-complete data instead of that of the censored data. As mentioned by Diebolt and Ip (1996), the outcome of such an algorithm, after a burn-in period, is a sequence of parameter values sampled from the stationary distribution of the algorithm. Its mean is close to the MLE result. Its dispersion reflects the information loss due to censoring.

3.4 BAYESIAN ESTIMATION (BE)

Now that the expectation step has been avoided, the tedious part of the SEM algorithm is the maximization step. It can also be avoided by putting the estimation problem into a Bayesian perpective. More precisely, assume that θ is a realization of a random parameter Θ with *prior* distribution p. Then the *posterior* distribution of Θ is

$$q(\theta|\ell_I, t_J) \propto p(\theta)r(\ell_I|\theta)[1 - R(t_J|\theta)]$$

The following algorithm has been designed so as to admit q for stationary distribution:
(i) generate $\theta \sim p$;
(ii) generate $\ell_J \sim r_\theta$;
(iii) generate $\theta \sim p(\cdot)r(\ell_I|\cdot)r(\ell_J|\cdot)$, and goto (ii).
This algorithm is nothing but a Gibbs sampler on (L_J, Θ). Step (ii) updates L_J while step (iii) updates Θ.

4 Weibull distribution

In order to implement MLE, EM, SEM and BE, an assumption must be made on an appropriate family of p.d.f. for g. Many choices are possible (Gamma, Pareto, Weibull etc...These distributions are described in full detail in Johnson and Kotz (1970)). In this paper, the actual trace lengths are supposed to follow a Weibull distribution with (unknown) parameter α and index b $(\alpha, b > 0)$

$$w_{\alpha,b}(\ell) = \alpha b \exp\left\{-(b\ell)^\alpha\right\}(b\ell)^{\alpha-1} \qquad \ell > 0 \tag{4}$$

If $L \sim w_{\alpha,b}$, then $bL \sim w_{\alpha,1}$. In other words, b is nothing but a scale factor. In contrast to this, the parameter α determines the shape of the distribution. If $\alpha < 1$, then $w_{\alpha,b}$ is monotonic decreasing and unbounded at the origin. If $\alpha > 1$, then

$w_{\alpha,b}$ is a unimodal distribution that is similar in shape to a normal distribution at large α values. In the intermediary case $\alpha = 1$, $w_{1,b}$ is an exponential distribution. The Weibull distribution has finite moments at all positive orders

$$E\big(L^n\big) = \frac{\Gamma\left(n\alpha^{-1}+1\right)}{b^n} \tag{5}$$

A Weibull distribution can be simulated either by inverting its distribution function or by considering a standard exponential variable U and then delivering $U^{1/\alpha}/b$.

The residual p.d.f. associated with the Weibull distribution is

$$r_{\alpha,b}(\ell) = \frac{b}{\Gamma(\alpha^{-1}+1)} \exp\left\{-(b\ell)^\alpha\right\} \qquad \ell > 0$$

This p.d.f. is monotonic decreasing whatever the values of α and b. The moments are equal to

$$E\big(R^n\big) = \frac{1}{b^n(n+1)} \frac{\Gamma\left((n+1)\alpha^{-1}+1\right)}{\Gamma\left(\alpha^{-1}+1\right)} \tag{6}$$

It can be noted that $E(R) < E(L)$ when $\alpha > 1$ as well as $E(R) > E(L)$ when $\alpha < 1$. The equality $E(R) = E(L)$ that takes place in the case $\alpha = 1$ stems from the lack of memory of the exponential distribution.
A simple way to generate a residual trace is to put $R = UV^{\frac{1}{\alpha}}$ where U an V are two independent variables respectively uniformly distributed on $]0,1[$ and gamma distributed with parameter $\alpha^{-1}+1$ and index b.

5 A simulation test

In order to test the efficiency of three of the procedures presented (MLE, SEM and BE), they have been applied to populations of residual traces sampled from $r_{0.5,1}$ with mean[3] $6m$ and variance $84m^2$. Each population can have 4 possible sizes $(100, 200, 500$ or 1000 traces), as well as 4 possible censoring levels $(0.924m, 2.817m, 7.250m$ and ∞, in accordance with the percentiles $75\%, 50\%, 25\%$ and 0%). For each of the $4 \times 4 = 16$ types considered, 10 populations have been simulated.

Figure 3 shows the influence of the size of the population, of the censorship proportion and of the type of estimator on the estimation of both parameters α and b. Several observations can be made:

1. The estimated points (α, b) are organized as elongated clouds;
2. those clouds tend to shorten as the size of the population increases;
3. the target point $(0.5, 1)$ is not offset;
4. the censorship proportion is mainly influential for large populations sizes;
5. the Bayesian clouds are shortest.

[3] To give a comparison, the mean and the variance of the Weibull distribution $w_{0.5,1}$ are respectively $2m$ and $20m^2$.

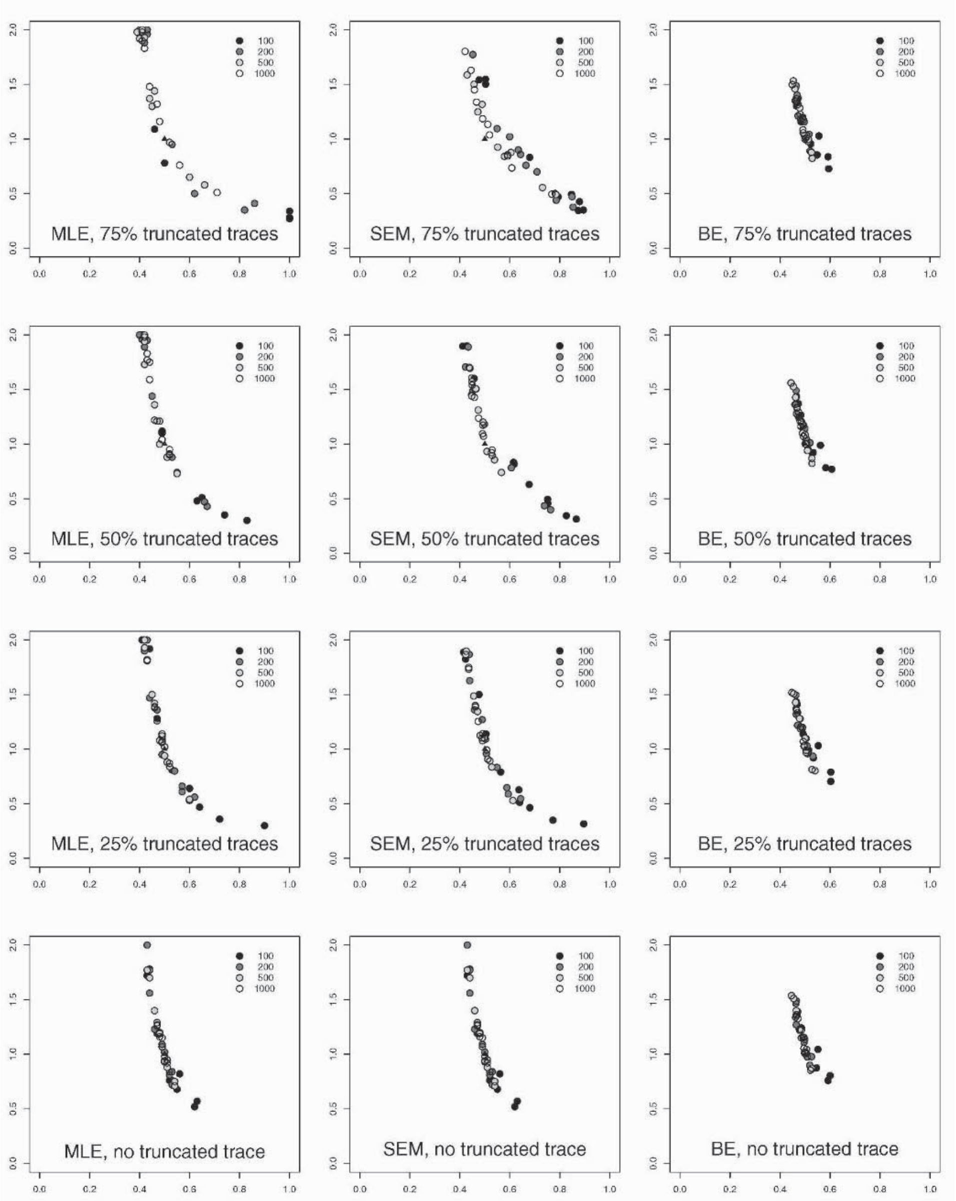


Figure 3. This figure plots the estimation of b versus that of α as a function of the estimator chosen, the number of traces in the population and the proportion of traces censored

The first observation suggests that both estimates $\hat{\alpha}$ and $\hat{b}$ are functionally dependent. To fix ideas, consider for instance the case of the MLE. By taking the partial derivative of the log-likelihood of the simulated data w.r.t. b, one can end up with an equation of the form

$$\hat{b} = f\left(\#I, \#J, \frac{\sum_{i\in I} \ell_i^{\hat{\alpha}}}{\#I}, t, \hat{\alpha}\right)$$

where f is a deterministic function, $\#I$ (resp. $\#J$) denotes the number of elements of I (resp. J) and t is the censoring level. In other words, the complete residual traces act in the estimation process only via their number and their empirical moments. In particular, if $\#I$ and $\#J$ have been fixed, all variability that can be expected in the parameter estimation derives from the statistical fluctuations of those empirical moments. When $\#I$ is large, they are not significantly different from the moment of order α of $r_{0.5,1}$ (see (6)) and the relationship between $\hat{\alpha}$ and $\hat{b}$ becomes deterministic.

The second observation is standard. The estimators have less and less variability as the population increases. In the case of large populations, the third observation indicates that the estimators tend to concentrate around the target point. In other words, the estimators are asymptotically unbiased.

The fourth observation is not surprising either. For large population sizes, the only factor that can affect the variability of the estimators is the censorship threshold.

The fifth observation suggests that the Bayesian procedure gives better results than MLE or SEM. This observation should be mitigated by the fact that the results obtained are highly dependent on the prior distribution chosen for (α, b). Here it has been supposed to be uniform over $]0,1[\times]0,2[$. If the range of uniformity of only one of the parameters had been extended, then the variability of both estimators would have been substantially increased.

6 Case study

The same three estimation procedures have been applied to a population of 419 traces taken from different outcrops embedded in the same geological formation, the Oxfordian limestones which overlie the Callovo-Oxfordian argilite formation of the underground research laboratory of ANDRA in the East of France. Fractures are subvertical and comprise faults and joints. A detailed structural and statistical study of the various fracture sets has been carried out (Bergerat *et al*, 2004). Here a directional set of subvertical joints is considered. The trace lengths range between $0.2m$ and $15m$ with a mean of $2.4m$ and a standard deviation of $2.5m$. Only 62 of the traces are censored (15%). Preliminary experiments suggested that one should certainly have $\alpha < 2$ as well as $b < 1$. This motivated us to apply the BE procedure with (α, b) *a priori* uniformly distributed on $]0,2[\times]0,1[$. On the other hand, the SEM and BE procedures have been resumed during 5000 iterations including a burn-in period of 1000 iterations. The 4000 pairs of values (α_n, b_n) generated by each procedure have been averaged to obtain the estimates of α and b of Table 1. This table gives also estimates of the mean and of the standard deviation of the

Weibull distribution. They have been obtained by calculating at first the mean m_n and the standard deviation σ_n associated to each (α_n, b_n), and then averaging them.

	$\hat{\alpha}$	$\hat{b}$	$\hat{m}$	$\hat{\sigma}$
MLE	0.860	0.370	2.919	3.407
SEM	0.837	0.464	2.366	2.841
BE	1.094	0.374	2.582	2.362

Table 1. Estimates obtained from the three procedures

The three estimated values for α and b cannot be considered as similar. Nonetheless, they give reasonably comparable estimations for the mean trace length. In contrast to this, the differences between the estimated standard deviations are more pronounced[4]

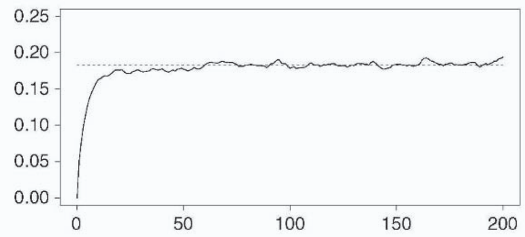


Figure 4. Variogram of the estimates of the mean along iterations

A potent advantage of Bayesian estimation is that it delivers a posterior distribution for the parameters under study, from which variances, quantiles as well as confidence limits can be deduced. For instance, it is possible to assign a variance to the estimate of the mean. As the values generated are dependent a simple approach is to consider the sill of the experimental variogram of the m_n's (see Fig. 4). We arrive at a variance of $0.183m^2$ (or a standard deviation of $0.43m$). Using similar approaches, it is also possible to attribute a variance to the standard deviation estimate ($0.050m^2$) or even a covariance between the mean and the standard deviation estimates ($0.048m^2$).

7 Discussion

In this paper, the traces have been considered as independent. This is of course a simplifying but not always appropriate assumption. In the case where joints tend to cluster or when they abut to the border of sedimentological layers, dependence relationships must be introduced between traces.

[4] It can also be mentioned that the trace length distribution was estimated in a previous exercise (Bergerat *et al*, 2004) using a MLE based on the gamma family. The estimated mean ($2.67m$) and the estimated standard deviation ($2.84m$) obtained are perfectly compatible with the results of this paper.

Another simplification has been made by assuming that the joints have their trace length independent of their orientation. If this assumption is not valid, Terzhagi's correction must be applied to compensate for the fact that a joint has more chance to be observable when its orientation is orthogonal to the outcrop (see Chilès and de Marsily (1993) and references therein).

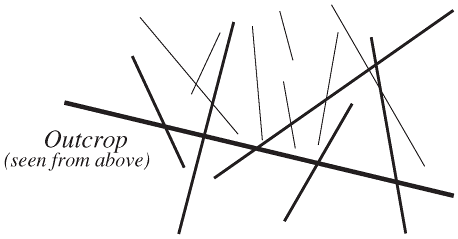


Figure 5. The more elongated the joint in the direction orthogonal to the outcrop, the more chance it has to be observed as a trace

One also may wonder what is the relationship between the trace length and the joint height distributions? To fix ideas, suppose that the joints are rectangles. Then a random joint has its statistical properties specified by the trivariate distribution of its width W, its height H and its dihedral angle Θ with the outcrop. The p.d.f. g of the trace lengths is related to that of the joint heights f by the formula

$$g(h) \propto f(h)E\{W \sin\Theta | H = h\}$$

Simplifications occur in the following cases:

1. If W and H are proportional, then $g(h) \propto hf(h)E\{\sin\Theta|H = h\}$;
2. If W and H are independent, then $g(h) \propto f(h)E\{\sin\Theta|H = h\}$;
3. If Θ is uniform on $]0, \pi/2[$, then $g(h) \propto f(h)E\{W|H = h\}$.

8 References

Bergerat F., Chilès J.P., Frizon de Lamotte D. and Elion P. (2004) - Analyse des microstructures tectoniques du Dogger et de l'Oxfordien calcaires. Bilan des Etudes et Travaux de l'Andra - 2003, chapitre II, fiche technique 2.4.2.

Chilès J.P. and de Marsily G. (1993) - Stochastic models of fracture systems and their use in flow and transport modeling. In Bear J., Tsang C.F. and de Marsily G. (eds) *Flow and contaminant transport in fractured rocks*, Academic Press, San Diego, pp. 169-236.

Dempster A.P., Laird N.M. and Rubin D.B. (1977) - Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B*, **39**, pp. 1-38.

Diebolt J. and Ip E.H.S. (1996) - Stochastic EM: method and application. In Gilks W.R., Richardson S. and Spiegelhalter D.J. (eds) *Markov chain Monte Carlo in practice*, CRC Press, Boca Raton, pp. 259-273.

Johnson N.L. and Kotz S. (1970) - Distributions in statistics (Volume 1). Wiley, New York.

Laslett G.M. (1982) - Censoring and edge effects in areal and line transect sampling of rocks joint traces. *Math. Geol.*, **14-2**, pp. 125-140.

ON SOME CONTROVERSIAL ISSUES OF GEOSTATISTICAL SIMULATION

L.Y. HU and M. LE RAVALEC-DUPIN
Department of Reservoir Engineering
Institut Français du Pétrole
1 et 4 avenue du Bois-Préau
92852 Rueil-Malmaison – France

Abstract. In this paper, we intend to clarify some conceptual issues of geostatistical simulation such as reproduction of the model covariance, equi-probability, independence, etc. We also introduce discussions on the confusion between simulating ergodic random functions and sampling random vectors. Our focus is on the interpretation of these probabilistic concepts in terms of realizations rather than the precision of simulation algorithms.

1 Introduction

Should conditional simulations reproduce the model covariance? This is one of the many controversial issues in geostatistics. Some argue that conditional and unconditional realizations are realizations of the same random function model and that they must reproduce the model covariance due to the ergodicity of the random function model. Of course, this reproduction is up to statistical fluctuations, i.e. fluctuations from the model parameters because of the limited size of a realization. Others argue that conditional realizations must respect the conditional (or posteriori) covariance but not the model (prior) covariance, and that this conditional covariance is different from the model prior covariance and even non-stationary whatever the model prior covariance.

Another famous controversial issue relates to the equi-probability of independently generated realizations. For ones, as random seed numbers are equi-probable, the resulting realizations are equi-probable too. For others, when in the Gaussian framework for instance, realizations (discretized as vectors) in the neighborhood of the mean vector are more likely to happen than the others. Therefore, realizations of a multi-Gaussian vector are not equi-probable even when independently generated.

The analysis of these contradictory points of view leads to other issues like

- Is there any (numerical) criterion to say that two (or several) realizations of a random function (in a large-enough domain) are independent?
- Can we generate an infinity of "independent" realizations of a random vector of finite dimension?

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 175-184.

- Is geostatistical simulation of random functions consistent with the sampling of multivariate distributions?

Understanding what is behind these issues is not only of philosophical interest, but also of great importance in the application of geostatistical methods to model calibration, model sampling and uncertainty estimation etc.

In this paper, we intend to clarify some of the above issues and to introduce discussions about some others. Our discussions are limited to the stationary (multi-)Gaussian random function. We focus on conceptual issues of numerical simulations rather than numerical precision of simulation algorithms. We also explore the significance of some well established concepts of probability (Feller, 1971) in terms of an individual realization or a set of realizations. We always assume that the simulation domain is large enough with respect to the covariance range.

2 Regional covariance and covariance matrix

2.1 REGIONAL COVARIANCE

We study physical properties that are unique and defined in a field. With the geostatistical approach, a physical property is considered as a realization of an ergodic random function. The ergodic property is necessary for the inference of the structural parameters (regional mean, variance and covariance etc.) of the random function model from a single realization. When a random function model is adopted to represent the physical property, we use the measurements (data) at some locations of the field to infer the structural parameters that specify the random function. Then we build realizations of the random function and each of these realizations should honor, up to statistical fluctuations, the inferred structural parameters due to ergodicity.

Assume that we have enough data to infer correctly the regional covariance. The uncertainty in the inference of the regional covariance is an important, but different issue. The reproduction of the regional covariance in geostatistical simulations is essential because this covariance is inferred from physical data and not just only prior idea. We believe that it is methodologically inconsistent to infer the covariance from a data set and then to build a realization, conditioned to the same data set, that has a covariance, i.e., the posterior covariance in the Bayesian terminology (Tarantola, 1987; de Marsily et al., 2001), conceptually different from the inferred one.

Let us examine how conditioning an unconditional realization by kriging preserves the regional covariance. Consider a stationary standard Gaussian random function $Y(x)$. Let $(Y(x_1),Y(x_2),...,Y(x_n))$ be a standard Gaussian vector and $Y^*(x)$ the simple kriging of $Y(x)$ using the covariance function $C(h)$ and the data set $(Y(x_1),Y(x_2),...,Y(x_n))$. Let $S(x)$ be a standard Gaussian random function with $C(h)$ as covariance function but independent of $Y(x)$, and $S^*(x)$ the simple kriging of $S(x)$ using the data set $(S(x_1),S(x_2),...,S(x_n))$. Then, $Y_c(x)$ defined by

$$Y_c(x) = Y^*(x) + S(x) - S^*(x)$$

is a standard Gaussian random function with $C(h)$ as covariance function, and $Y_c(x)$ is conditioned to the random vector $(Y(x_1), Y(x_2), ..., Y(x_n))$.

We note that the proof of the above result in Journel and Huijbregts (1978) or in Chilès and Delfiner (1999) assumes that the conditioning data set is a random vector. When the conditioning data set is fixed, $Y_c(x)$ becomes actually non-stationary. In particular, the mean values of $Y_c(x)$ at the data locations $x_1, x_2, ..., x_n$ equal respectively the data values, and the variances of $Y_c(x)$ at these locations are zero. In general, the covariance of the random function $Y_c(x)$ with a fixed conditioning data set is non-stationary (dependent on the location x) and therefore different from $C(h)$.

However, the fact that the covariance of the random function $Y_c(x)$ with a fixed conditioning data set is non-stationary does not necessarily mean that the regional covariance of an individual realization of $Y_c(x)$ would not reproduce the model covariance $C(h)$. Indeed, because $Y_c(x)$, conditioned to the random vector $(Y(x_1), Y(x_2), ..., Y(x_n))$, is a stationary ergodic random function, all realizations of $Y_c(x)$ should reproduce the covariance function $C(h)$ up to statistical fluctuations. Consequently, for any fixed data set $(y(x_1), y(x_2), ..., y(x_n))$, i.e. a realization of $(Y(x_1), Y(x_2), ..., Y(x_n))$, and any realization $s(x)$ of $S(x)$, $y_c(x)$ defined by

$$y_c(x) = y^*(x) + s(x) - s^*(x)$$

is a Gaussian realization conditioned to $(y(x_1), y(x_2), ..., y(x_n))$ and provides $C(h)$ as its regional covariance up to statistical fluctuations.

The difference and the relation between the regional covariance and the covariance of a random function should become clearer by examining the concept of covariance matrix.

2.2 COVARIANCE MATRIX

In practice, we often need to discretize a random function over a finite grid. So we deal with random vectors and we can define their covariance matrixes. In the literature, the covariance matrix in the Bayesian framework is related to a set of realizations, instead of a single realization. For instance, the posterior covariance matrix of a random vector (after being conditioned to a data set) is related to the set of conditional realizations (not to a single conditional realization). Similarly, the prior covariance matrix of a random vector (before being conditioned to a data set) is related to the set of unconditional realizations (not to a single unconditional realization).

It is important not to confound the regional covariance of an individual realization with the covariance matrix of an ensemble of realizations. Any conditional simulation of an ergodic random function (discretized over a grid) must preserve the regional covariance but not the prior covariance matrix that will certainly change after conditioning.

The above discussion can be further clarified through the following example. Let $(y(x_1), y(x_2),..., y(x_N))$ be an unconditional realization of the Gaussian random vector $(Y(x_1), Y(x_2),..., Y(x_N))$. We use, for instance, the sequential Gaussian simulation method for generating realizations and we assume that all conditional distributions are computed without any approximation. When N is large enough and when the grid nodes $(x_1, x_2,..., x_N)$ covers a domain much larger than the area delimited by the covariance range, the experimental (regional) covariance should reproduce the theoretical covariance. Now, consider $y(x_1)$ as a conditioning datum, and we generate realizations of $(Y(x_2),..., Y(x_N))$ conditioned to $y(x_1)$. By using the same random numbers for sampling the conditional distributions at the nodes $(x_2,..., x_N)$ as in the case of the above unconditional simulation, we obtain the realization $(y(x_1), y(x_2),..., y(x_N))$ conditioned to $y(x_1)$. This conditional realization is identical to the unconditional realization and therefore has the same experimental (regional) covariance. However, if we generate a set of realizations conditioned to $y(x_1)$, their covariance matrix will be different from the model prior covariance.

In general, an unconditional realization $(y(x_1), y(x_2),..., y(x_N))$ of a random vector $(Y(x_1), Y(x_2),..., Y(x_N))$ can always be seen as a realization conditioned to $(y(x_1), y(x_2),..., y(x_I))$ for $I < N$. Evidently, this suggests that the conditioning does not necessarily change the regional covariance.

2.3 SUMMARY

The regional covariance and the covariance matrix (in the Bayesian terminology) are two different concepts in geostatistical simulation. A conditional simulation method should guaranty that the regional covariance of each conditional realization reproduces, up to statistical fluctuation, the model covariance function. However, the covariance matrix of a set of conditional realizations is conceptually (not because of statistical fluctuations) different from that of a set of unconditional realizations.

The covariance reproduction in geostatistical simulation means the reproduction of the regional covariance, not the (prior) covariance matrix. The reproduction of a covariance matrix is a much stronger requirement than that of a regional covariance. Reproduction of a covariance matrix requires generating a large-enough number of realizations that represent correctly the multivariate probability distribution, while the regional covariance is related to a single realization.

Because of the uniqueness of the physical property under study, the regional covariance has a physical sense, while the covariance matrix is a model concept.

3 Equi-probability and likelihood

"Are realizations of a stochastic model equi-probable?" is another controversial issue that still troubles practitioners of geostatistics. Considering a set of realizations as equi-probable or not can change completely the way we evaluate uncertainties from these realizations.

3.1 EQUI-PROBABILITY

Consider, for instance, the numerical simulation of a stationary Gaussian random function of order 2 over a finite grid of the simulation field. Namely, we simulate a Gaussian vector Y of N components $(Y(x_1), Y(x_2),..., Y(x_N))$. Now if we generate K realizations of Y: $y_1, y_2,..., y_K$, starting from K independent uniform numbers (random seeds issued from a random number generator), these realizations are equi-probable. This is because the uniform seeds can be considered as equi-probable, and for a given seed, a simulation algorithm produces a unique realization of Y after a series of deterministic operations.

3.2 LIKELIHOOD

However, the probability density values of the random vector Y at $y_1, y_2,..., y_K$ are different. Consider two realizations y_1 and y_2 and assume that $g(y_1) > g(y_2)$, where g stands for the probability density function of Y. Thus, we are more likely to generate realizations in the neighborhood of y_1 than in that of y_2. In other words, for a given small-enough domain $\delta(y)$ located at y and a large-enough number of realizations of the vector Y, there are more realizations in $\delta(y_1)$ than in $\delta(y_2)$. But this does not mean y_1 is more probable to happen than y_2. Consequently, when evaluating uncertainty using a set of independently generated realizations, they must be equally considered with the same weight.

3.3 SUMMARY

Before generating realizations, there is a larger probability to generate a realization in the neighborhood of the realization of higher probability density. But once a set of realizations is generated independently between each other, they are all equi-probable.

4 Independence, correlation and orthogonality

4.1 INDEPENDENCE

When performing numerical simulation of the random function model, one often needs to generate more than one realization. These realizations are said independent because they are built by using the same simulation procedure but with different random seeds. These random seeds are said statistically independent. This is meaningful when a large number of random seeds are generated. Now, if we generate only a few realizations, say only two realizations, we need then only two random seeds. Because it does not make sense to talk about statistical independence with only two fixed numbers, does it make sense to talk about independence of two realizations of a random function model?

But geostatisticians are used to build models with only two "independent" realizations. This is the case when building realizations of the intrinsic model of coregionalization (Matheron, 1965; Chilès and Delfiner, 1999), when perturbing a realization by substituting some of its values with some other "independent" values (Oliver et al., 1997), when performing a combination of two independent realizations within the gradual deformation method (Hu, 2000), when modifying a realization using the probability perturbation method (Caers, 2002), etc.

If it does not make sense to check the independence between two realizations, it is nevertheless meaningful, at least for large realizations, to evaluate the degree of their correlation.

4.2 CORRELATION

Consider again the N-dimensional standard Gaussian vector $Y=(Y(x_1),Y(x_2),...,Y(x_N))$. Let $y_i=(y_i(x_1),y_i(x_2),...,y_i(x_N))$ $(i=1,2,...,I)$ be I independent realizations of Y. For each realization y_i, we compute its mean and its variance:

$$m_i=\frac{1}{N}\sum_{n=1}^{N}y_i(x_n)$$

$$\sigma_i^2=\frac{1}{N}\sum_{n=1}^{N}\left[y_i(x_n)-m_i\right]^2$$

When N is large enough and when the grid $(x_1,x_2,...,x_N)$ covers a domain whose dimension in any direction is much larger than the covariance range, we have $m_i\approx 0$, $\sigma_i^2\approx 1$. For any two realizations y_i and y_j, we usually compute their correlation coefficient as follows:

$$r_{ij}=\frac{1}{N}\sum_{n=1}^{N}\left[\frac{y_i(x_n)-m_i}{\sigma_i}\right]\left[\frac{y_j(x_n)-m_j}{\sigma_j}\right]$$

We have $r_{ij} = 1$ for $i = j$, and because of the "independence" between y_i and y_j for $i \neq j$, we have $r_{ij} \approx 0$. In practice, if the correlation coefficient r_{ij} $(i \neq j)$ is significantly different from zero, then the two realizations y_i and y_j are considered as correlated, and therefore dependent.

4.3 ORTHOGONALITY

Now for any two realizations y_i and y_j, we define the following inner product:

$$\left\langle y_i, y_j \right\rangle = \frac{1}{N}\sum_{n=1}^{N} y_i(x_n) y_j(x_n)$$

We have $\left\langle y_i, y_j \right\rangle \approx r_{ij} \approx 0$ for $i \neq j$ and $\left\langle y_i, y_j \right\rangle \approx r_{ij} = 1$ for $i = j$. Therefore, when $I = N$, the N vectors $y_i = (y_i(x_1), y_i(x_2),..., y_i(x_N))$ $(i = 1,2,..., N)$ constitute an orthonormal basis (up to statistical fluctuations) of an N-dimensional vector space V_N furnished with the above inner product. All other realizations of the random vector Y can be written as linear combinations of these N independent realizations $y_1, y_2,..., y_N$. In other words, we cannot generate more than N realizations of an N-dimensional random vector so that the above usual correlation coefficient between any two of these realizations equals zero.

4.4 CONSEQUENCE

The above remark has an unfortunate consequence for many iterative methods that involve the successive use of independent realizations. For instance, the gradual deformation method requires generating, at each iteration, a realization independent from all realizations generated at previous iterations. When the number of iterations of the gradual deformation method becomes equal to or larger than the number of grid nodes, the optimized realization at iteration l $(l \geq N)$ and a new realization at iteration $l+1$ are linearly dependent (not because of statistical fluctuations). Consequently, the condition for applying the gradual deformation method with combination of independent realizations is no longer satisfied when the number of iterations is larger than the number of grid nodes. This explains why when applying the gradual deformation method with a large number of iterations (much larger than the number of grid nodes), it is possible to progressively force the optimized realization to have a regional covariance different from the initial one (Le Ravalec-Dupin and Noetinger, 2002).

Nevertheless, in practice, the number of iterations is hopefully much smaller than the number of grid nodes. Otherwise, the method is not applicable when the calculation of the objective function requires heavy computing resources.

5 Random function or random vector?

Up to now, we have assumed that an ergodic random function can be represented by a random vector. This seems questionable. Consider the following experimentation of thought. Let $y(x)$ be an ergodic realization of the standard Gaussian random function $Y(x)$. Assume now $k(x) = \exp[y(x)]$ represents a physical property distributed in a field, say rock permeability in an oil field. The "real" permeability field is then completely known. Starting from a large-enough data set of $k(x)$, we can infer the covariance of the random function model $Y(x)$.

Now because we generate realizations over a finite grid $(x_1, x_2, ..., x_N)$, we deal with a Gaussian vector $Y = (Y(x_1), Y(x_2), ..., Y(x_N))$. As discussed before, there is a non-negative probability to generate realizations in a given domain $\delta(y)$ located at y. For a domain of fixed size, this probability is maximal when y equals the mean vector. If we sample correctly the random vector Y, there is a non-negative probability to generate realizations in the neighborhood of the mean vector. These realizations have small regional variances (smaller than the model variance 1) and their regional covariance will not respect the model covariance inferred from the "reality": $y(x) = \ln[k(x)]$!

The above reasoning (if it makes sense) leads to the following consequences:

- Geostatistical simulations (over a finite grid) cannot honor both the regional covariance and the multivariate probability density function.

- The sequential simulation algorithm (Johnson, 1987; Deutsch and Journel, 1992) is related to random vectors, and therefore it cannot generate realizations of ergodic random functions (Lantuéjoul, 2002), even in the Gaussian case where we can compute the conditional distribution without approximation by using the global neighborhood.

- The use of an exact sampling method based on the Markov iteration or the acceptation/rejection (Omre, 2000) will make it possible to generate a set of realizations representative of the multivariate probability density function. Due to the maximum likelihood of the mean vector, we must expect some realizations close to the smooth mean vector. This is not compatible with the foundation of geostatistics whose aim is to model spatial variability inferred from real data.

6 Conclusions and further discussions

If the theory of geostatistics based on simulating ergodic random functions is compatible with that based on sampling random vectors, then we have the following conclusions:

- The regional covariance of each conditional or unconditional simulation should reproduce, up to statistical fluctuations, the model covariance function inferred from real data.
- A large-enough set of conditional (or unconditional) simulations should respect, up to statistical fluctuations, the conditional (or unconditional) covariance matrix.
- A set of independently generated realizations are equi-probable but can have different probability density values.
- We cannot generate more then N orthogonal realizations of an N-dimensional random vector, if we use the usual correlation coefficient as the measure of correlation between realizations.

However, it seems that the theory based on the exact sampling of random vectors is contradictory with that based on the simulation of ergodic random functions. If this is true, there are then two possible theories of geostatistics: one based on random functions and the other based on random vectors. In the framework of the random function based geostatistics, we can talk about ergodicity, regional covariance and its inference from a single (fragmentary) realization (i.e., the real data set). In the framework of the random vector based geostatistics, we can talk about covariance matrix, but not ergodicity (that is not defined). The inference of the model parameters from a single realization is then questionable. These two theories are self-consistent and but they seem not compatible between each other. To avoid, at least, terminological confusion, it is necessary to choose one of the two frameworks: random vectors or random functions. In practice, we should expect that these two theories converge to each other with huge random vectors and random functions in large field.

Note finally that, in most real situations, the primary concern in geostatistical modeling remains the choice of a physically realistic random function (or set) model. For instance, a multi-Gaussian model is in contradiction with many geological settings such as fluvial channel or fractured system (Gomez-Hernandez, 1997). Then comes the difficulty of building realizations that preserve the spatial statistics inferred from data and that are calibrated to all quantitative (static and dynamic) data. The further evaluation of uncertainty is meaningful only under the following conditions: first the probability density function (pdf), conditioned to all quantitative data, covers correctly the range of uncertainty and second enough samples (realizations) of this conditional pdf can be obtained within an affordable time. The second condition depends on the efficiency of the sampling algorithms and the computer resources. But the first condition depends on the degree of objectivity of the pdf model. Because of the uniqueness of the reservoir property of interest, a pdf model should largely be subjective. We can evaluate uncertainty only within an subjective model (Matheron, 1978; Journel, 1994). The preservation of the model spatial statistics and the model calibration to data are objective problems, while the uncertainty evaluation is a subjective one (although mathematically meaningful and challenging within a pdf model).

References

Caers, J., 2002, Geostatistical history matching under training-image based geological model constraints, Paper SPE 77429.

Chilès, J.P. and Delfiner, P.,1999, Geostatistics - Modeling spatial uncertainty, Wiley, New York, 695p.

Deutsch, C.V. and Journel, A.G., 1992, GSLIB - Geostatistical software library and user's guide, Oxford University Press, New York, 340p.

Feller, W., 1971, An introduction to probability theory and its applications, Vol.I and Vol.II, John Wiley & Sons, New York, 509p. and 669p.

Gomez-Hernandez, J.J., 1997, Issues on environmental risk assessment, In E.Y. Baafi and N.A. Schofield (eds.), Geostatistics Wollongong '96, Kluwer Academic Pub., Dordrecht, Vol.1, p.15-26.

Hu, L.Y., 2000, Gradual deformation and iterative calibration of Gaussian-related stochastic models, Math. Geology, Vol.32, No.1, p.87-108.

Johnson, M., 1987, Multivariate statistical simulation, John Wiley & Sons, New York, 230p.

Journel, A.G. and Huijbregts, C. J., 1978, Mining geostatistics, Academic Press, London, 600p.

Journel, A.G., 1994, Modeling uncertainty: some conceptual thoughts, In Dimitrakopoulos (ed.), Geostatistics for the next century, Kluwer Academic Pub., Dordrecht, p.30-43.

Lantuéjoul, C., 2002, Geostatistical simulation - Models and algorithms, Springer-Verlag, Berlin, 256p.

Le Ravalec-Dupin, M., and Noetinger, B., 2002, Optimization with the gradual deformation method, Math. Geology, Vol.34, No.2, p.125-142.

Marsily, G. de, Delhomme, J.P., Coudrain-Ribstein, A., Lavenue, A.M., 2001, Four decades of inverse problems in hydrogeology, In Zhang and Winter (eds.), Theory, Modeling and Field Investigation in Hydrogeology, Geophysical Society of America, Special Paper 348, p.1-17.

Matheron, G., 1965, Les variables régionalisées et leur estimation - Une application de la théorie des fonctions aléatoires aux sciences de la nature, Masson, Paris, 305p.

Matheron, G., 1978, Estimer et choisir, Fascicule 7, Les Cahiers du Centre de Morphologie Mathématique de Fontainebleau, Ecole des Mines de Paris, 175p.

Oliver, D.S., Cunha, L.B. and Reynolds, A.C., 1997, Markov chain Monte Carlo methods for conditioning a permeability field to pressure data, Math. Geology, vol. 29, no. 1, p. 61-91.

Omre, H., 2000, Stochastic reservoir models conditioned to non-linear production history observations, In W. Kleingeld and D. Krige (eds.), Geostatistics 2000 Cape Town, Vol.1, p.166-175.

Tarantola, A., 1987, Inverse problem theory - Methods for data fitting and model parameter estimation, Elsevier, Amsterdam, 613p.

ON THE AUTOMATIC INFERENCE AND MODELLING OF A SET OF INDICATOR COVARIANCES AND CROSS-COVARIANCES

EULOGIO PARDO-IGÚZQUIZA[1] and PETER A. DOWD[2]
[1]*Department of Mining and Mineral Engineering,*
University of Leeds, Leeds LS2 9JT, UK
[2]*Faculty of Engineering, Computer and Mathematical Sciences,*
University of Adelaide, Adelaide, SA 5005, Australia

Abstract. The indicator approach to estimating spatial, local cumulative distributions is a well-known, non-parametric alternative to classical linear (ordinary kriging) and non-linear (disjunctive kriging) geostatistics approaches. The advantages of the method are that it is distribution-free and non-parametric, is capable of dealing with data with very skewed distributions, provides a complete solution to the estimation problem and accounts for high connectivity of extreme values. The main drawback associated with the procedure is the amount of inference required. For example, if the distribution function is defined by 15 discrete thresholds, then 15 indicator covariances and 105 indicator cross-covariances must be estimated and models fitted. Simplifications, such as median indicator kriging, have been introduced to address this problem rather than using the theoretically preferable indicator cokriging. In this paper we propose a method in which the inference and modelling of a complete set of indicator covariances and cross-covariances is done automatically in an efficient and flexible manner. The inference is simplified by using relationships derived for indicators in which the indicator cross-covariances are written in terms of the direct indicator covariances. The procedure has been implemented in a public domain computer program the use of which is illustrated by a case study. This technique facilitates the use of the full indicator approach instead of the various simplified alternatives.

1 Introduction

The general estimation problem can be stated as the estimation at unsampled locations of the most probable value of a variable together with a measure of the uncertainty of the estimation (e.g. estimation variance). A more complete and interesting solution to the problem, however, is to estimate at each unsampled location the local cumulative distribution function (cdf) conditioned to the neighbouring data. Point estimates, interval estimates, measures of uncertainty and probabilities (e.g. probability of the unknown value being greater than a specified threshold) can be easily obtained from the estimated distribution function,. The indicator approach (Journel, 1983; Goovaerts, 1997) offers a non-parametric solution to the problem of estimating such local cdf. A discrete representation of the cdf is defined by K thresholds from which K indicator random functions can be derived by applying the thresholds to the continuous variable $Z(u)$:

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 185-193.

$$I(u;z_k) \quad = \quad \begin{cases} 1 & \text{if } Z(u) \leq z_k \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

Each indicator random function is assumed to be second-order stationary, i.e. unbounded semi-variograms are not allowed and covariances and semi-variograms are equivalent statistical tools. The expected values of the indicator variables are interpreted as the distribution function of the continuous variable for the respective thresholds:

$$\mathrm{E}\{I(u;z_k\} \;=\; F(u;z_k) \;=\; \Pr\{Z(u) \leq z_k\} \tag{2}$$

Estimating the different indicators provides estimates of the cdf, $F(u;z_k)$, for the different thresholds $\{z_k; k=1,...,K\}$. The complete cdf is estimated by assuming a form of the cdf between the thresholds and for the tails (Deustch and Journel, 1992; Goovaerts, 1997).

Journel and Alabert (1989) argue that indicator cokriging is theoretically the best estimator (in a least squares sense) of the cdf using indicators from the experimental data for all the thresholds simultaneously. This is, however, an onerous procedure requiring the inference of K^2 indicator covariances and cross-covariances. In practice, the cross-covariances are assumed to be symmetric and the number of models is reduced to K indicator covariances and $K(K\text{-}1)/2$ indicator cross-covariances. An asymmetrical cross-covariance implies (from the non-centred cross-covariance) that:

$$\mathrm{P}\{Z(u) \leq z_k, Z(u+h) \leq z_{k'}\} \neq \mathrm{P}\{Z(u) \leq z_{k'}, Z(u+h) \leq z_k\} \tag{3}$$

and:

$$\mathrm{P}\{Z(u) \leq z_k\} \neq \mathrm{P}\{Z(u+h) \leq z_k\} \tag{4}$$

If the indicator random functions are second-order stationary, the random function $Z(u)$ is distribution-ergodic (Papoulis, 1984), the restriction in (4) no longer holds and thus symmetrical cross-covariances are justified.

Even with the assumption of symmetric cross-covariances with, say, $K = 10$ there are 10 indicator covariances and 45 indicator cross-covariances to infer and model. The inference of direct indicator covariances is not particularly difficult and can be done more or less automatically by using maximum likelihood (Pardo-Igúzquiza, 1998) to infer the parameters (e.g., range, sill, nugget, anisotropy angle) without the need to estimate the covariance for a number of lags and fit a model. Even using maximum likelihood the inference and modelling of indicator cross-covariances is more difficult because, *inter alia*, of the order relations (Journel and Posa, 1990) which impose restrictions on the types of models that can be fitted to the indicator cross-covariances. A much more efficient and routine procedure would be to express the cross-covariances in terms of the direct indicator covariances.

2 Methodology

Given a continuous variable, $Z(u)$, and its cdf, the range of values of the variable is represented by K discrete thresholds. Given any pair of these thresholds, z_k and $z_{k'}$, (with the convention $z_{k'} > z_k$) a class, or categorical, variable, c_k, can be defined with an associated indicator random function $I(u;c_k)$:

$$I\left(u;c_k\right) \; = \; \begin{cases} 1 & \text{if } Z\left(u\right) \in c_k \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

or, equivalently:

$$I\left(u;c_k\right) \; = \; \begin{cases} 1 & \text{if } z_k < Z\left(u\right) \le z_{k'} \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

with

$$I(u;c_k) = I(u;z_{k'}) - I(u,z_k) \tag{7}$$

The covariance of the indicator random function for the class on the left hand side must be equal to the covariance of the difference between the indicator random functions for the thresholds on the right hand side:

$$\text{Cov}\{I(u;c_k)\} = \text{Cov}\{I(u;z_{k'}) - I(u,z_k)\} \tag{8}$$

or $$\text{Cov}\{I(u;c_k)\} = \text{Cov}\{I(u;z_{k'})\} + \text{Cov}\{I(u,z_k)\} - 2\text{Cov}\{I(u;z_k), I(u;z_{k'})\} \tag{9}$$

then $$C_I(h;z_k,z_{k'}) = \frac{1}{2}\left\{C_I(h;z_{k'}) \; + C_I(h;z_k) \; - \; C_I(h;c_k)\right\} \tag{10}$$

which expresses the indicator cross-covariance of each pair of thresholds as a function of the direct indicator covariances at the thresholds and for the class that they define. There are *K*(*K*+1)/2 models, all defined by indicator covariances, which can be efficiently modelled by maximum likelihood and where:

$$C_I(h;z_k,z_{k'}) = \text{Cov}\left\{I(u;z_k), I(u+h;z_{k'})\right\} = \text{E}\left\{I(u;z_k)I(u+h;z_{k'})\right\} - F(z_k)F(z_{k'}) \tag{11}$$

$$C_I(h;z_k) = \text{Cov}\left\{I(u;z_k)\,I(u+h;z_k)\right\} = \text{E}\left\{I(u;z_k)I(u+h;z_k)\right\} - F^2(z_k) \tag{12}$$

$$C_I(h;c_k) = \text{Cov}\left\{I(u;c_k)\,I(u+h;c_k)\right\} = \text{E}\{I(u;c_k)I(u+h;c_k)\} - \left(F(z_{k'}) - F(z_k)\right)^2 \tag{13}$$

Note that, in general, Equation (10) defines a composite model for the indicator cross-covariance even if the direct indicator covariances are simple models. There is no need to fit a specific model to the indicator cross-covariance as it is defined by the indicator covariances and Equation (10).

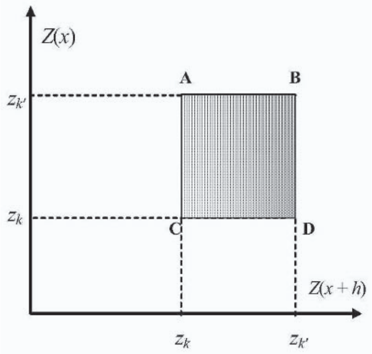


Fig. 1. General order relation

Journel and Posa (1990) give the order relations for indicator covariances and indicator cross-covariances. The order relations follow from the fact that the non-centred indicator covariances $K_I(h;z_k,z_{k'})$ are bivariate cumulative distribution functions, with

$$K_I(h;z_k,z_{k'}) = C_I(h;z_k,z_{k'}) + F(z_k)F(z_{k'}) \tag{14}$$

The general order relation can be written as (Journel and Posa, 1990):

$$C_I(h;z_k)+C_I(h;z_{k'})+E\geq 2C_I(h;z_k,z_{k'}) \qquad (15)$$

with $E=F^2(z_k)+F^2(z_{k'})-2F(z_k)F(z_{k'})$. This order relation is easily derived from Figure 1 in which the shaded area represents a bivariate cumulative distribution function, i.e. a probability which must be non-negative. Then $B-A-D+C\geq 0$, where each of the corners is a non-centred indicator covariance and

$$K_I(h;z_{k'})-K_I(h;z_k,z_{k'})-K_I(h;z_{k'},z_k)+K_I(h;z_k)\geq 0$$

from which (15) can be easily derived taking account of (14) and the assumption of a distribution-ergodic random function *Z*(*u*). Thus the non-centred indicator cross-covariance is symmetric with respect to the thresholds.

Substituting (10) into (15) gives:

$$C_I(h;c_k)\geq 2F(z_k)F(z_{k'})-F^2(z_k)-F^2(z_{k'}) \qquad (16)$$

which can be written as:

$$C_I(h;c_k)\geq -\left(F(z_k)-F(z_{k'})\right)^2 \qquad (17)$$

As the term on the right hand side is always negative this inequality will be satisfied if the covariance of the class indicator is positive. This inequality shows that (10) conforms to the general order relation given by (15).

3 Case study

A realization of a non-Gaussian random function was generated on an 80 × 80 grid by sequential indicator simulation using the program sisimm (Deutsch and Journel, 1992). The thresholds, quantiles and covariance models used for generating the realization are given in Table 1 and a plot of the realization is shown in Figure 2. The range increases as the quantile increases, implying that there is greater connectivity of high values than low values - the semi-variogram range for the 0.9 quantile is four times larger than that for the symmetrical quantile with respect to the median (0.1).

Threshold number	z_k	$F(z_k)$	Range of isotropic spherical covariance model (length units)
1	0.1	0.1	6
2	0.5	0.3	8
3	2.5	0.5	10
4	5.0	0.7	12
5	10.0	0.9	24

Table 1. Thresholds and indicator models used in generating the realization shown in Figure 2 using sisim (Deutsch and Journel, 1992).

A sample of 60 randomly located data was drawn from the realisation; the values of the samples are represented in Figure 3. Maximum likelihood inference does not require *Z*(*u*) to be Gaussian; in fact in this application it is applied to binary indicator data. Nevertheless, using the likelihood of the multivariate normal distribution provides an efficient estimator of the indicator covariance parameters (Pardo-Igúzquiza, 1998). The procedure is also useful for model selection as shown hereafter.

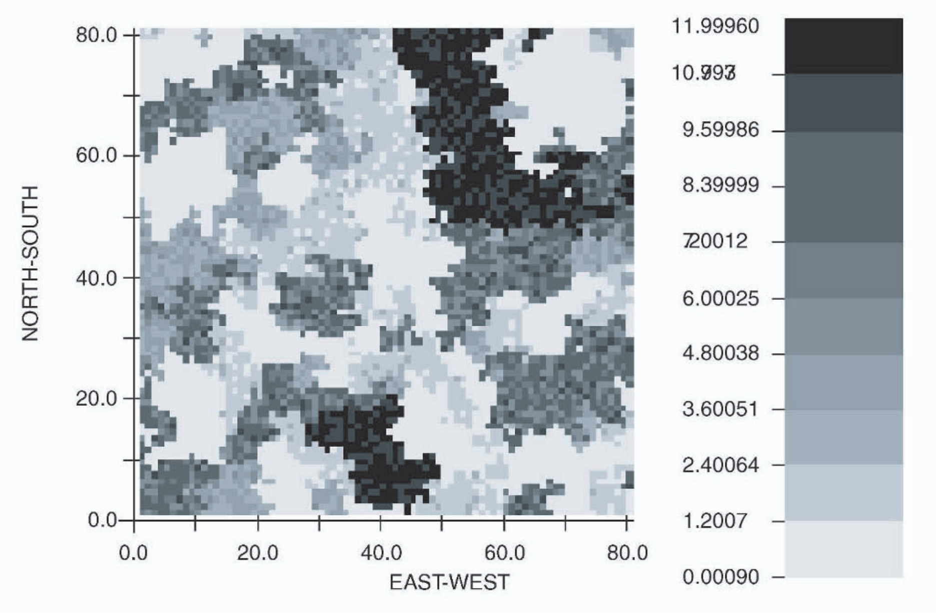


Fig. 2. Simulated random field using indicator sequential simulation with the parameters given in Table 1.

For a specified type of model (spherical, exponential or Gaussian) the program infers the model parameters of the indicator covariance for each threshold and all classes. Eight models were assessed: (1) one isotropic model with no nugget; (2) one isotropic model with nugget; (3) one anisotropic model with no nugget; (4) one anisotropic model with nugget; (5) two nested isotropic models with no nugget; (6) two nested isotropic models with nugget; (7) two nested models: one isotropic, one anisotropic and no nugget; (8) two nested models: one isotropic, one anisotropic and a nugget.

The number of parameters ranges from two for model (0) to seven for model (7) - nugget, sill of isotropic model, range of is otropic model, sill of anisotropic model, long range, short range and anisotropy angle. The most appropriate model could simply be chosen by inspection of the method of moments estimate of the semi-variogram (Figure 4) for the direct indicator semi-variograms at the thresholds. Any of the models can be tried and the quality of the fit can be assessed by the value of the

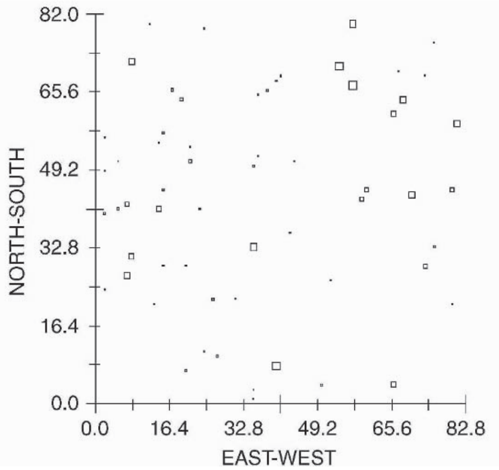


Fig. 3 Pictogram of 60 sample values with locations selected at random from the 80 × 80 grid of the simulated field.

negative log-likelihood function (NLLF). However, the model becomes more flexible as more parameters are used, and a lower NLLF may be achieved by a meaningless over-specification. A model selection criterion, such as the Akaike information criterion (AIC) (Akaike, 1974), provides a trade-off between simple models and more exact fits:

$$AIC(\ell) = 2L(\ell) + 2k(\ell)$$

where: $AIC(\ell)$ is the Akaike information criterion value for the ℓ-th model,

$L(\ell)$ is the value of the negative log-likelihood function for the ℓ-th model, and

$k(\ell)$ is the number of independent parameters fitted in the ℓ-th model.

The model with the lowest $AIC(\ell)$ value is chosen.

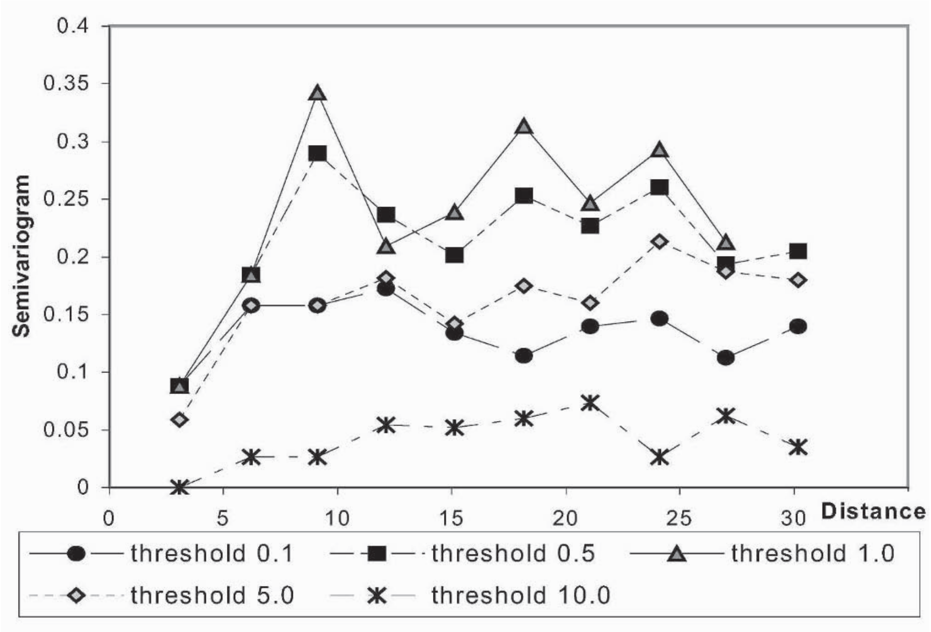


Fig. 4 Experimental indicator semi-variograms for the five thresholds

For few observations simple models should be chosen as, in general, there is insufficient evidence in the data for a model with a large number of parameters. In such cases using a large number of parameters amounts to modelling the fluctuations generated by sampling variability.

Table 2 shows the spherical model fitted by the program using the 60 observations shown in Figure 3 for model 1. In terms of the AIC values the best model is one isotropic structure with no nugget, which is the model used to generate the simulated realization. When comparing estimated parameters with those used in the simulation it should be remembered that only 60 randomly located data were used for the estimates and, as a consequence, they are subject to a high degree of sampling variability. Nevertheless, the ranges of the spherical models are quite well estimated.

Indicator covariances

Threshold		Covariance parameters				Anisotropy	Model
I	AIC	C_0	C	Range 1	Range 2	angle	type
1	57.440	0.0	0.149	5.883	5.883	0.0	1
2	78.399	0.0	0.226	8.818	8.818	0.0	1
3	77.549	0.0	0.226	9.336	9.336	0.0	1
4	63.744	0.0	0.204	13.996	13.996	0.0	1
5	-27.800	0.0	0.058	22.627	22.627	0.0	1

Indicator covariances for the indicator classes

Threshold			covariance parameters				Anisotropy	Model
I_1	I_2	AIC	C_0	C	Range 1	Range 2	angle	type
1	2	44.997	0.0	0.136	6.228	6.228	0.0	1
1	3	79.625	0.0	0.249	8.473	8.473	0.0	1
1	4	83.529	0.0	0.249	12.098	12.098	0.0	1
1	5	68.098	0.0	0.185	11.752	11.752	0.0	1
2	3	64.314	0.0	0.214	7.782	7.782	0.0	1
2	4	80.712	0.0	0.233	6.401	6.401	0.0	1
2	5	82.137	0.0	0.241	9.336	9.336	0.0	1
3	4	26.942	0.0	0.056	10.371	10.371	0.0	1
3	5	74.151	0.0	0.202	9.336	9.336	0.0	1
4	5	55.506	0.0	0.173	13.996	13.996	0.0	1

Table 2. Results for one isotropic structure and no nugget, i.e. two parameters: variance (sill) and range. I is indicator number, model type 1 is spherical.

From Table 2 and using (10) the models shown in Figure 5 are fitted to the indicator cross-covariances. The example has been restricted to five thresholds to limit the size of tables and number of figures, but even for large numbers of thresholds the procedure is computationally efficient, e.g. the program generates the 120 models for 15 thresholds in a few minutes.

4 Conclusions

The indicator cokriging of local cumulative distributions is often avoided because of the burden of modelling a large number of indicator covariances and cross-covariances. The authors have described a procedure that bases the modelling of the indicator cross-covariances on direct models of indicator covariances for the thresholds and for the classes defined by pairs of thresholds. The direct indicator covariances can be efficiently inferred and modelled by maximum likelihood which can be applied without assuming that the continuous random variable is multivariate Gaussian.

A public domain program, available on request from the authors, allows the modelling of a wide range of structures, with or without nugget, isotropic or anisotropic, and with up to two nested models. Each structure may be spherical, exponential or Gaussian. The AIC, provided by the program for each indicator covariance, can be used for model

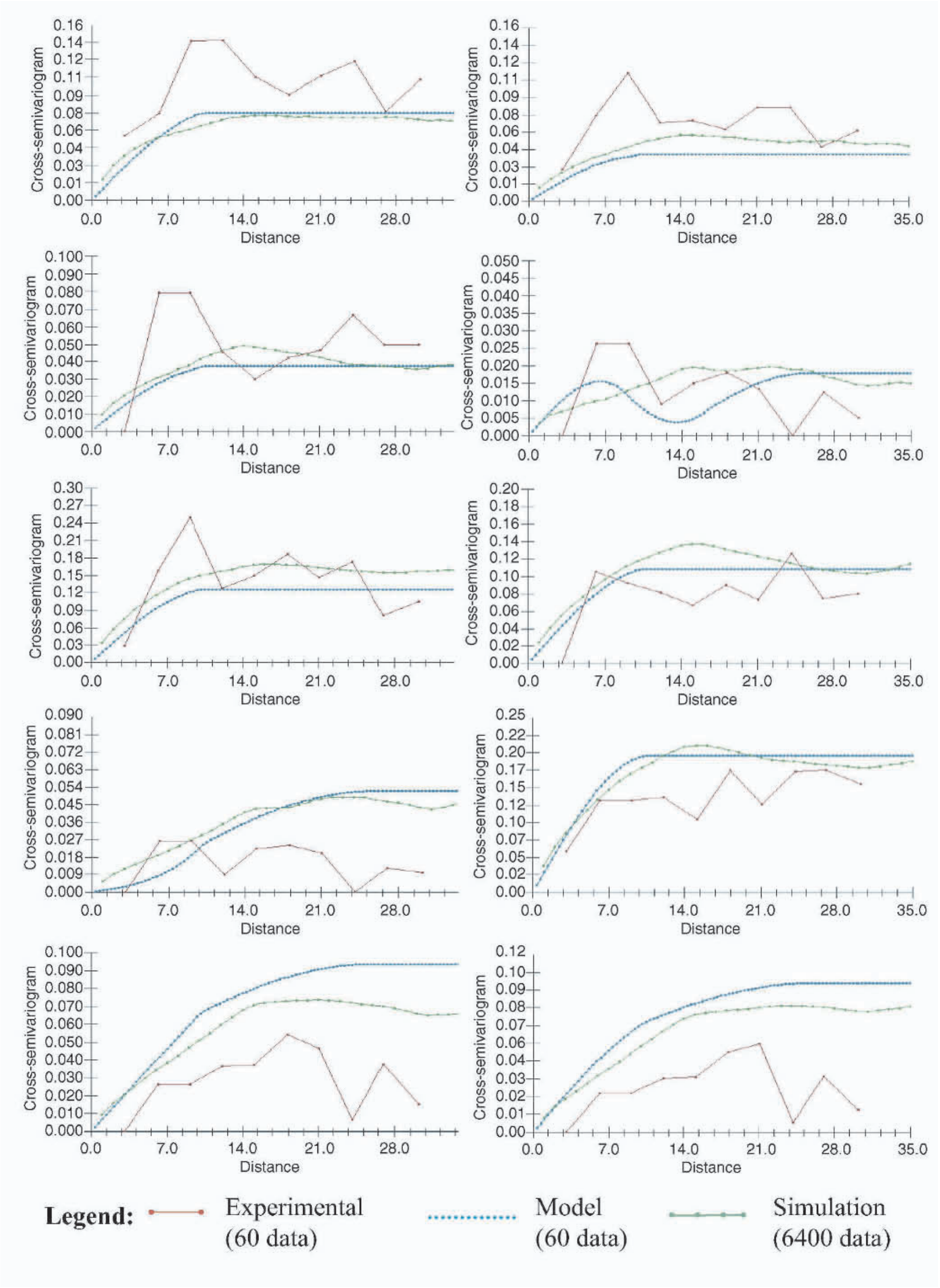


Fig. 5 Indicator cross-semivariograms and models fitted by (10) and the maximum likelihood estimates given in Table 2. Left to right and top to bottom: (a) thresholds 1 and 2; (b) thresholds 1 and 3; (c) thresholds 1 and 4; (d) thresholds 1 and 5; (e) thresholds 2 and 3; (f) thresholds 2 and 4; (g) thresholds 2 and 5; (h) thresholds 3 and 4; (i) thresholds 3 and 5; (j) thresholds 4 and 5.

selection. A case study illustrated the methodology on a simulated realization of a random field.

Acknowledgements

The work reported in this paper was funded by UK Engineering and Physical Sciences Research Council Grant number GR/M72944.

References

Akaike, H., A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, vol. AC-19, no. 6, 1974, p. 716-723.

Deutsch, C.V. and Journel, A.G., GSLIB. Geostatistical Software Library and User's Guide. Oxford University Press, New York, 1992, 340 p.

Goovaerts, P., Geostatistics for Natural Resources Evaluation. Oxford University Press, New York, 1997, 483p.

Journel, A. G., Nonparametric estimation of spatial distributions. *Mathematical Geology*, vol. 15 no. 3, 1983, p. 445-468.

Journel, A. G. and Alabert, F., Non-Gaussian data expansion in the Earth Sciences. *Terra Nova*, vol. 1, 1989, p. 123-134.

Journel, A. G. and Posa, D., 1990. Characteristic behavior and order relations for indicator variograms. *Mathematical Geology*, vol. 22, no. 8, 1990, p. 1011-1025.

Papoulis, A., Probability, Random Variables and Stochastic Processes. McGraw-Hill International Editions, Singapore, 1984, 576 p.

Pardo-Igúzquiza, E., Inference of spatial indicator covariance parameters by maximum likelihood using MLREML. *Computers and Geosciences*, vol. 24, no. 5, 1998, p. 453-464.

This page intentionally blank

ON SIMPLIFICATIONS OF COKRIGING

JACQUES RIVOIRARD

Centre de Géostatistique, Ecole des Mines de Paris, Fontainebleau - France

Abstract. Due to the number of variables or of data, cokriging can be a heavy operation, requiring simplifications. Two basic types of simplications, with no loss of information, are considered in this comprehensive paper. The first type of simplifications consists, in the isotopic case, in reducing cokriging to kriging, either of one or several target variables, or of spatially uncorrelated factors. The example of variables linked by a closure relation (e.g. constant sum, such as the indicators of disjoint sets) is in particular considered. The other type of simplifications is related to some particular models that, in given configurations, screen out a possibly large part of data. This results in simplified and various types of heterotopic neighborhoods, such as collocated, dislocated, or transferred.

1 Introduction

Given a set of multivariate data, cokriging theoretically improves the linear estimation of one variable by taking into account the other variables. In particular, it ensures consistency between the estimates of different variables: the cokriging of a linear combination of variables equals the linear combination of their cokriging, which is in general not the case for kriging. For instance, the cokriged estimate of the difference between the top and the bottom elevations of a geological layer is the same as the difference of their cokriged estimates. Similarly, if we consider sets that correspond for instance to classes of values of a random function, cokriging ensures the relation:

$$[1_{Z(x)<z_2}]^{CK} = [1_{Z(x)<z_1}]^{CK} + [1_{z_1 \leq Z(x)<z_2}]^{CK}$$

which is clearly desirable. In addition cokriging is central in multivariable simulation in a Gaussian framework, where regressions are linear and where the absence of correlation is equivalent to independence.

However, due to the ever increasing number of data (in term of data points or of measured variables), cokriging can become rapidly a heavy operation, hence the interest in looking for simplifications, either in the isotopic cases (when all variables are known at each data point), or in the heterotopic cases. In this comprehensive paper, two types of simplifications without loss of information (i.e. the estimation coinciding with full cokriging) are considered: isotopic cases where cokriging of some variable reduces to

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 195-203.

its kriging (other variables being screened out), and cokriging neighbourhood being simplified by screening out some types of data.

In the following, we will consider *p* variables Z$_1$(x), …, Z$_i$(x), …, Z$_p$(x). For simplicity the variables are supposed second order stationary with simple and cross-covariances:

$$C_i(h) = C_{ii}(h) = Cov\left[Z_i(x), Z_i(x+h)\right]$$
$$C_{ij}(h) = Cov\left[Z_i(x), Z_j(x+h)\right]$$

or (in particular when cross-covariances are even functions) intrinsic with simple and cross-variograms:

$$\gamma_i(h) = \gamma_{ii}(h) = \frac{1}{2}E\left(\left[Z_i(x+h) - Z_i(x)\right]^2\right)$$
$$\gamma_{ij}(h) = \frac{1}{2}E\left[Z_i(x+h) - Z_i(x)\right]\left[Z_j(x+h) - Z_j(x)\right]$$

The difficult problem of estimating consistently multivariate structures when all variables are not known at all data points is not addressed in this paper. Note that different choices may be possible for the variables (e.g. the indicators of disjoint classes of values of a random function, or the indicators of accumulated classes), that are theoretically equivalent. However one choice may be found to be easier to detect possible simplifications of cokriging.

2 Reduction of cokriging to kriging in isotopic cases

2.1 SELF-KRIGEABILITY

Given a set of *p* variables ($p \geq 2$), one of the variables, for instance Z$_1$, is said to be self-krigeable, if its cokriging coincides with its own kriging in any isotopic configuration (Matheron 1979, Wackernagel 2003). The necessary and sufficient condition for this is its cross-structure with the other variables being identical (more exactly, proportional) to its own structure, which is denoted (a little abusively when the proportionality factor is zero) as:

$$C_{1j}(h) \equiv C_1(h)$$

or

$$\gamma_{1j}(h) \equiv \gamma_1(h)$$

for all *j*, and can possibly be checked using sample simple and cross-structures. (Note that the concept of self-krigeability is relative to a given set of variables: for instance a self-krigeable variable may not remain self-krigeable if new variables are added.)

Suppose now that, among a set of p variables ($p \geq 2$), two of them are self-krigeable, say Z$_1$ and Z$_2$: $\gamma_{12}(h) \equiv \gamma_1(h)$ and $\gamma_{12}(h) \equiv \gamma_2(h)$. Then two cases can be distinguished. Either they have the same structure (the cross-structure being proportional to these, not excluding it being zero):

$$\gamma_{12}(h) \equiv \gamma_1(h) \equiv \gamma_2(h)$$

in which case they are intrinsically correlated, in the sense that the correlation coefficient between Z$_1$(v) and Z$_2$(v) within a domain V is "intrinsic", not depending on the support v nor on the domain V (Matheron, 1965; this should not to be confused with the intrinsic model based on increments). Or they have different structures, in which case the variables (or their increments in the intrinsic model) are necessarily spatially uncorrelated:

$$\gamma_{12}(h) = 0 \quad \forall h$$

Suppose now that all p variables are self-krigeable. We can group the variables that have the same simple structure (up to a proportionality factor). Then all variables within a given group are intrinsically correlated. Moreover two variables from different groups are spatially uncorrelated. So a set of self-krigeable variables can be partitioned into groups of intrinsically correlated variables, each group having a different structure, and with no correlation between groups (Rivoirard, 2003; another proof is given in Subramanyam and Pandalai, 2004). In particular we have the two typical cases of reduction of cokriging to kriging:
- when all variables are intrinsically correlated, all simple and cross-structures being proportional to a common structure, and kriging weights being the same for all variables;
- when they have no cross correlation.

In the particular case of p self-krigeable variables being linked by a closure relation (such as concentrations, or such as the indicators of disjoint sets):

$$Z_1(x) + Z_2(x) + ... + Z_p(x) = 1$$

or more generally being linearly dependent, separating the different groups yields:

$$\text{var}[\sum Z_i(x)] = \text{var}[\sum_{group} \sum_{i \in group} Z_i(x)] = \sum_{group} \text{var}[\sum_{i \in group} Z_i(x)] = 0 \Rightarrow \text{var}[\sum_{i \in group} Z_i(x)] = 0 \quad \text{for}$$

each group.

Hence each group of intrinsically correlated variables is closed. Note that, in the case of the indicators of disjoint sets partitioning the space, there can be only one group, since these indicators are necessarily correlated. So if the indicators of disjoint sets are self-krigeable, their cross and simple structures are proportional to a unique structure (which corresponds to the mosaic model with independent valuations, Matheron (1982)).

Linear dependency between variables is not a problem for defining cokriging (e.g. as a projection). The cokriged value is perfectly determined, but in the isotopic case, the cokriging system is singular and weights are not all uniquely defined. So it is interesting to remove (at least) one of the variables. Providing that linear dependency vanishes, cokriging does not depend on the choice of which variables are removed. However some choices may be better than others, in term of simplification of cokriging. Suppose that *p*-1 out of a set of *p* variables are self-krigeable, say Z_1, …, Z_i, …, Z_{p-1}. If these *p*-1 variables are intrinsically correlated, then all *p* variables are intrinsically correlated and for each variable, cokriging reduces to kriging of that variable. But suppose now that there is more than one group of intrinsically correlated variables in the *p*-1 variables (which cannot be the case for the indicators of disjoint sets). Then the last variable:

$$Z_p = 1 - Z_1 - Z_2 - \ldots - Z_{p-1}$$

which is necessarily correlated to at least some of the *p*-1 variables since:

$$\operatorname{var} Z_p = -\sum_{i<p} \operatorname{cov}(Z_p, Z_i) > 0$$

cannot be self-krigeable: as groups are uncorrelated, the cross-structure with each group is proportional to the structure of the group, and so these cross-structures cannot all be identical (they are either different, or possibly equal to zero but only for some of them). Then cokriging is simplified by kriging the *p*-1 self-krigeable variables, and deducing the estimation of the last one.

2.2 FACTORIZATION

Up to now, we have considered simplification of isotopic cokriging resulting from initial variables being self-krigeable, which can be detected directly from the observation of simple and cross-structures. We will consider now an extension through the use of factors.

2.2.1 Model with residual

If we consider a set of $p = 2$ variables Z_1 and Z_2, with Z_1 being self-krigeable:

$$\gamma_{12}(h) = a\ \gamma_1(h)$$

the residual of the linear regression of $Z_2(x)$ on $Z_1(x)$ at same point x:

$$R(x) = Z_2(x) - a\ Z_1(x) - b$$

(by construction with mean zero and uncorrelated to $Z_1(x)$ at same point *x*) has no spatial correlation with Z_1. In this model of residual, or "Markov" type model (Journel, 1999; Chilès and Delfiner 1999; Rivoirard, 2001), the variable Z_2 is subordinated to the master variable Z_1:

$$Z_2(x) = a\, Z_1(x) + b + R(x)$$

The model is factorized into Z_1 and R, which (being spatially uncorrelated) are self-krigeable. So we have:

$$Z_1^{CK} = Z_1^K$$

$$R^{CK} = R^K$$

$$Z_2^{CK} = a\, Z_1^K + b + R^K$$

for any isotopic configuration. This model was illustrated in a mining case study by Bordessoule et al. (1989): this included the deduction of the self-krigeability of a variable from the observed simple and cross-variograms, the analysis of the residual, and the cokriging, showing in particular that cokriging can be significantly different from kriging in practice, if there was any doubt.

Note that if the structure of the residual is identical to this of Z_1, all simple and cross-structures are identical, so that Z_1 and Z_2 are intrinsically correlated. Then any variable can be taken as master.

2.2.2 Intrinsically correlated variables

If variables have simple and cross-structures proportional to a common structure (say a correlogram $\rho(h)$), so do their linear combinations. So intrinsic correlation between a set of variables extends to their linear combinations. It follows that if two variables, or two linear combinations of variables, are uncorrelated at same point ($C_{ij}(0) = 0$), their cross-covariance is identically zero:

$$C_{ij}(h) = C_{ij}(0)\rho(h) = 0$$

So, for intrinsically correlated variables, the absence of statistical correlation implies the absence of geostatistical, or spatial, correlation. As a consequence, any statistical factorization (eigen vectors, successive residuals, etc) of intrinsically correlated variables gives spatially uncorrelated factors (Rivoirard, 2003). Note that, while arbitrary or conventional in the sense they depend on the choice of the factorization method, these factors are objective, in the sense their values at a point where the variables are known, are determined. While factorisation is always possible when the model is admissible (exhibiting factors ensures the variances-covariances matrix to be valid), it does not provide simplification in isotopic cokriging, as all variables and linear combinations are self-krigeable anyway.

2.2.3 Linear model of coregionalization

The linear model of coregionalization corresponds to a decomposition of structures into a set of basic structural components (e.g. describing different scales). The corresponding components of the variables for a given scale are intrinsically correlated (Journel and Huijbregts, 1978), and a factorisation of these ensures the validity of the model. However there is usually more factors than variables, and these factors have not an objective meaning. Factorial (co-)kriging, or kriging analysis, allows estimating consistently these, but does not simplify cokriging.

2.2.4 Objective factors

Cokriging can be greatly simplified when the variables are factorized into objective factors, for the knowledge of the variables at a data point is equivalent to the knowledge of factors. And since the factors are spatially uncorrelated, they are self-krigeable, yielding cokriging of all linear combinations, and in particular of the initial variables. Note that some factors can share the same structure (in which case these factors are intrinsically correlated and their choice is conventional).

Of course a question is how to determine, if possible, such factors in practice. In general a statistical factorization (non-correlation at same point) does not yield zero cross-correlation. The technique of min-max autocorrelation factors (Desbarats and Dimitrakopoulos, 2000; Switzer and Green, 1984) allows building factors that are not only uncorrelated at distance 0, but also at a distance chosen from sampling: then the lack of correlation must be assumed or checked for other distances. This reduces cokriging to kriging of factors, and in the gaussian case, multivariate simulation to separate simulations of factors. Another approach, seeking the absence of correlation simultaneously for all lags of variograms, is proposed by Xie and Myers (1995) and Xie et al. (1995).

The absence of cross-correlation between factors for all distances also corresponds to the isofactorial models of non-linear geostatistics (disjunctive kriging, i.e. cokriging of indicators, being obtained by kriging the factors).

3 Simplifications of cokriging neighbourhood in heterotopic cases

3.1 DATA EXPRESSED AS INDIVIDUAL VALUES OF VARIABLES OR FACTORS

Simplifications coming from non spatially correlated variables are still valid in heterotopic cases, when data consist of individual values of these variables. For instance, if Z_1 and Z_2 are spatially uncorrelated, Z_2 is screened out when cokriging Z_1, whatever the configuration. An exception must be noted, when uncorrelated variables have means, or drifts, that are unknown but related (Helterbrand and Cressie, 1994). The screen is deleted by the estimation of means which is implicitly performed within cokriging. We do not consider this case in this paper. Screen can also be deleted by data that do not consist of individual values of variables. If Z_1 and Z_2 are spatially uncorrelated for instance, the knowledge of e.g. the sole sum $Z_1 + Z_2$ at a data point can delete the screen.

Similarly, the simplifications due to factorized models hold in heterotopic cases where data consist in values of individual factors. In the model of residual, with Z_2 subordinated to master variable Z_1, factors are Z_1 and the residual, and simplifications hold providing that data can be equivalently expressed in term of Z_1 and Z_2, or of Z_1 and R. This occurs when Z_1 is known at each data point, for the possibly additional knowledge of Z_2 at this point is equivalent to that of the residual.

More generally the cokriging of a self-krigeable variable reduces to kriging of that variable, in heterotopic cases where it is known at every data point (Helterbrand and Cressie, 1994), for the possibly available other variables at these points are screened out. Of course it is not necessarily so in other heterotopic cases (this is why the definition of a self-krigeable variable assumes isotopy).

If the target variable is Z_2, subordinated to one (or possibly several) master variable Z_1 informed at all data points, its cokriging at any target point is simplified:

$$Z_2^{CK} = a\, Z_1^{K} + b + R^{K}$$

where Z_1 is kriged from all data points and R from only the data points where $Z_{2,}$ or equivalently R, is known. If Z_1 is known at any desired point, so in particular at target point, we have:

$$Z_2^{CK} = a\, Z_1 \ + b + R^{K}$$

and cokriging reduces to kriging of the residual (Rivoirard, 2001). Then cokriging is collocated, making use of the auxiliary variable only at target point and at points where the target variable is known, not at other data points. In other models, the residual R is spatially correlated to the auxiliary variable Z_1. Then, the auxiliary variable Z_1 still being supposed known at all desired points, the advantage of collocated cokriging is to be more precise (in term of estimation variance) than kriging the residual, for it corresponds to cokrige this residual from the same data. However, by using a collocated neighbourhood, both collocated cokriging and kriging of the residual result in a loss of information compared to full cokriging, when the residual is spatially correlated to the auxiliary variable, i.e. when the cross-structure is not proportional to this of the auxiliary variable.

3.2 OTHER SIMPLIFICATIONS OF NEIGHBOURHOOD

Consider for instance the case where the cross-structure is proportional to that of the target variable, not of the auxiliary variable, i.e. the target variable is the master variable Z_1. In the case the residual is pure nugget, knowing additionally Z_2 at a Z_1 data point corresponds to knowing the residual. Being uncorrelated to all data and to the target, this is screened out in any Simple Cokriging configuration, and the neighbourhood is dislocated, making use of the auxiliary variable only at target point (if available) and where Z_1 is unknown. In Ordinary Cokriging, the Z_2 values where Z_1 is known are not screened out for they participate to the estimation of the Z_2 mean, but all receive the same weight, which also simplifies cokriging.

Such simplifications of the cokriging neighbourhood are studied in detail by Rivoirard (2004), where a number of other simplifications are listed (with possible extensions to more than two variables). In particular (assuming target is unknown):

- If the target variable is master and has a pure nugget structure, it is screened out from Simple Cokriging where known alone, in the case the auxiliary variable is available at target point; else all data are screened out.
- If the target variable is subordinated to an auxiliary nugget and master variable, this last is screened out where known alone, except at target point if available, in any Simple Cokriging configuration.
- If the target variable is subordinated to an auxiliary master variable with a nugget residual, it is screened out from Simple Cokriging where the auxiliary variable is known (neighbourhood being transferred to the auxiliary variable for common data points); if additionally the auxiliary variable is available at target point, all data except this are screened out from Simple Cokriging.

4 Conclusions

In this paper, different simplifications of cokriging have been finally considered:

- cokriging reduced to kriging for a self-krigeable variable, in isotopic cases or when it is known at all data points;
- cokriging obtained from the kriging of spatially uncorrelated variables or factors (e.g. residuals), in isotopic cases or when data consist of individual values for these factors;
- screening out of some type of data in the neighbourhood, in some particular models with residual where master variable or residual are pure nugget.

Some of these simplifications, in given configurations, can be directly deduced from the observation of the simple and cross-structures of the variables (self-krigeability, intrinsic correlation, model with residual). Other simplifications depend on the possibility of building objective factors that are not cross-correlated. An open question is how to measure the efficiency of the simplifications for a possible departure from the assumptions.

References

Bordessoule, J. L., Demange, C., and Rivoirard, J., 1989, Using an orthogonal residual between ore and metal to estimate in-situ resources, *in* M. Armstrong, ed., Geostatistics: Kluwer, Dordrecht, v. 2, p. 923-934.

Chilès, J.-P., and Delfiner, P., 1999, Geostatistics: Modeling spatial uncertainty, Wiley, New York, 695 p.

Desbarats, A. J. and Dimitrakopoulos, R., 2000, Geostatistical simulation of regionalized pore-size distributions using min/max autocorrelation factors. *Math. Geol*. Vol. 32 No 8, p. 919-942.

Helterbrand, J. D., and Cressie, N., 1994, Universal cokriging under intrinsic coregionalization, Math. Geol.ogy, v. 26, no. 2, p. 205-226.

Journel, A.G. and Huijbregts, C.J., *Mining Geostatistics*, Academic Press, 1978.

Journel, A., 1999, Markov models for cross-covariances: Math. Geology, v. 31, no. 8, p. 955-964.

Matheron, G., 1965, Les variables régionalisées et leur estimation, Masson, Paris, 306 p.

Matheron, G., 1979, Recherche de simplification dans un problème de cokrigeage, Technical Report N-628, Centre de Géostatistique, Fontainebleau, France, 19 p.

Matheron, G., 1982, La destruction des hautes teneurs et le krigeage des indicatrices, Technical Report N-761, Centre de Géostatistique, Fontainebleau, France, 33 p.

Rivoirard, J., 2001, Which Models for Collocated Cokriging?: Math. Geology, v.33, no. 2, p. 117-131.

Rivoirard, J., 2003, Course on multivariate geostatistics, C-173, Centre de Géostatistique, Fontainebleau, France, 76 p.

Rivoirard, J., 2004, On some simplifications of cokriging neighborhood: Math. Geology, v.36, no. 8, p. 899-915.

Subramanyam A. and Pandalai, H. S., 2004, On the equivalence of the cokriging and kriging systems, Math. Geology, v. 36, no. 4.

Switzer, P. and Green, A. A., 1984, Min/max autocorrelation factors for multivariate spatial imaging. Technical report no 6, Department of Statistics, Stanford University, 14 p.

Wackernagel, H., 2003, Multivariate geostatistics: an introduction with applications, 3rd ed., Springer, Berlin, 387 p.

Xie, T. and Myers, D.E., 1995. Fitting matrix-valued variogram models by simultaneous diagonalization (Part I: Theory), Math. Geology, v. 27, p. 867-876.

Xie, T., Myers, D.E. and Long, A.E. 1995, Fitting matrix-valued variogram models by simultaneous diagonalization (Part I: Application), Math. Geology, v. 27, p. 877-888.

This page intentionally blank

EFFICIENT SIMULATION TECHNIQUES FOR UNCERTAINTY QUANTIFICATION ON CONTINUOUS VARIABLES:
A process preserving the bounds, including uncertainty on data, uncertainty on average, and local uncertainty

BIVER P.
TOTAL SA
Centre Scientifique et Technique Jean Feger
Avenue Larribau
64000 Pau, France

Abstract. In the petroleum industry, the standard Monte Carlo technique applied on global parameters (rock volume, average petrophysics) is often used to evaluate hydrocarbon in place uncertainty. With the increasing power of computers, these methodologies have become old fashioned compared to geostatistics. However care must be taken in using the latter blindly; multiple geostatistical realisations do not cover a reasonable uncertainty domain because of undesirable effects. For instance: a net to gross map with a small variogram range tends to reproduce the prior mean over the domain of interest even if this prior mean is established on a small data set; uncertainty on the data themselves may produce values outside the prior distribution range; more over the alternative data set may be biased comparing to initial prior distribution.
In this paper, we present a fully automatic process based on the classical normal score transformation which is able to handle, in a non-stationary model, the following characteristics:

- uncertainty on the data set (with systematic biases)
- uncertainty on the prior mean
- local uncertainty
- preservation of bounds defined on the prior model

An example is given on a real field case in the framework of net to gross modelling. The beta law is used in order to provide high frequencies observed at the bounds (0,1); the robustness of an automatic fit for this distribution type is highlighted in order to adjust a non-stationary model on the data set. All aspects described above have been handled successfully in this non-stationary context; and the ensemble of realisations reproduces rigorously the prior distribution. The balance between local uncertainty and global uncertainty is provided by the user; consequently the volumetrics distribution are easily controlled. A final comparison with the classical geostatistical workflow is provided.

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 205-214.

1 Introduction

A continuous variable has to be depicted with a complete probability density function; it could be a non parametric density function issued from smoothed histogram (when a large number of data are available) or a parametric model fitted to the noisy experimental histogram; for instance: a porosity histogram is often calibrated with a Gaussian model, but for a net to gross histogram the beta distribution can be more appropriate.

Except for the Gaussian model, the assumption of kriging requires a normal score transform to process the data and build a random field; the back-transform is subsequently applied to come back to the real variable.

The motivation in the framework of multiple realisations is to incorporate the uncertainty on data and the uncertainty on the mean; and at the same time to retrieve the probability density function over all the possible realisations especially in the case of bounded models (uniform, triangular and beta distribution). This goal is achieved with the use of the normal score transformation, in a generalized context, coupled with a kriging of the mean.

2 Uncertainty on the mean

The uncertainty evaluation exercise may be performed in the appraisal phase of a reservoir; at this stage sparse data are available and it is not obvious that the prior mean (derived from data analysis) is perfectly well known. More frequently, this mean is uncertain and this uncertainty has a direct impact on hydrocarbon in place distribution. The quantification of the relative uncertainty on the mean can be assessed with a declustering formula.

Let us define, for the variable to simulate *Z(x)*,

a relative dispersion of the mean in the real space $r = \sigma_m / \sigma_t = 1/\sqrt{n}$

with - σ_m the uncertainty on the mean in the real space

- σ_t the global uncertainty on *Z(x)*

- *n* the number of independent data

Let us assume a covariance model *C(h)* for *Z(x)*; *n* can be derived from the ordinary kriging system if the covariance vector is set to zero (ordinary kriging of the mean *m*);

$$m = \sum_{i=1}^{n} \lambda_i . Z(x_i) \tag{1}$$

we have : $\sum_{j=1}^{n} C_{ij} . \lambda_j = \mu$

$$\sum_{i=1}^{n} \lambda_i = 1$$

with λ_i the kriging weights and μ the Lagrange parameter
It can be shown that: $\sigma_{OK}^2 = \sigma_m^2 = C(0).\mu$ hence $\mu = 1/n$ as $C(0) = \sigma_t^2$ (2)

If we draw a new mean value in the real space and if we want to preserve the distribution shape of the residuals, it is necessary to shift the distribution; as a consequence, the bounding values are not preserved from one realisation to another. An alternative procedure is to work in the normal score space and to consider that the standard normal distribution is in fact a sum of two random variables:

- the first random variable Δ characterises the uncertainty of the mean; it has an infinite spatial correlation, it is centred on zero and has a standard deviation of $r = \sqrt{\mu} = 1/\sqrt{n} = \sigma_m / \sigma_t$ (relative dispersion of the mean in the real space);
- the second random variable characterises the uncertainty of the residuals; it has a limited spatial correlation, it is centred on Δ and has a standard deviation of $\sigma_r = \sqrt{1 - \frac{1}{n}}$ (3)

Each realisation (characterised by its Δ) is then back transformed with the global cumulative distribution of the variable. As a consequence, each realisation has a specific distribution according to the value of Δ; it is different from the initial global distribution. However, if we consider the ensemble of all these different realisations, the global distribution is retrieved. Despite of this property, we have to check that the uncertainty on the mean in the real space (induced by the variation of Δ) is consistent with the relative dispersion we want to impose.

The procedure is illustrated on Figure 1. Two initial global distributions are considered; a beta distribution between 0 and 1 with shape parameters p = 0.7 and q = 0.4 (this model could be appropriate to represent the distribution of a net to gross) and a Gaussian distribution with mean m = 0.2 and standard deviation σ = 0.06 (this model could represent a porosity uncertainty). The number *n* of equivalent independent data is set to 4 (*r = 0.5*). The updating of the distribution for each value of Δ is represented on Figure 1; for the beta distribution, the asymmetry is gradually modified with Δ ; for the Gaussian case, the shape is preserved and the influence of Δ is only a resizing of the Gaussian curve.

By construction, the global distribution is perfectly reproduced. The histograms of the mean values in real space are depicted on Figure 2. Concerning the mean, the distribution in the Gaussian case is of course an exact reproduction of the reference; the distribution in the beta case is slightly skewed comparing to the reference. Even with this high value of *r = 0.5*, the approximation of the mean dispersion is excellent.

If more asymmetric distributions are envisioned, the uncertainty on the mean could be asymmetric for high values of r but when r is decreased, it converges rapidly to the reference Gaussian distribution; this property is a consequence of the central limit theorem. Concerning the variance, the drawback of considering the variation of the mean in the normal score domain is that the variance is not constant for all realisations but perfectly correlated to the mean value.

The methodology can be extended in non-stationary cases. In these situations, the global cumulative distribution is substituted by a local cumulative distribution and, as a consequence, the normal score transform is locally adapted; however, in the Gaussian space, the simulation strategy with Δ is unchanged and still stationary. Its effect in real space is different, the uncertainty on the mean will be higher in the location where local cumulative distribution corresponds to a larger dispersion. In this case, the influence of the variable mean and variable trend dip (that can be major sources of uncertainty, see Massonnat, Biver, Poujol) are considered in one single step.

3 Uncertainty on the data

The data used in practical applications are not always considered as "hard" data. Today, log analysts are able to quantify uncertainties on their data; two kinds of uncertainties can be considered:

- measurements uncertainties linked to the resolution of logging tools,
- interpretation uncertainties linked to the parameters of the interpretation law chosen to derive interpreted logs from raw logs (for instance the exponents of the Archie's laws or the resistivity of connate water used to derive water saturation).

The first is mainly a noise on the data set, the second introduces a systematic bias from one data set to another; they are treated differently in the data set simulation procedure handled by the log analysts.

However, if systematic biases are suspected, an up date of the distribution is needed for each data set. Let's assume that we have multiple realizations of the data set of interest, and that a histogram and a prior distribution model are derived from the ensemble of data set realizations. The suggested procedure of updating can be described as following:

- compute normal score transform $G(xi)$ of the current data set realization x_i $i=1,n$
- compute the potential bias distribution in the normal score space with ordinary kriging of the mean for $G(x)$, using previously mentioned declustering formula with the normal score transform variogram model ;
- draw a value of the bias Δ' in its distribution with mean $m_{\Delta'}$ and standard deviation $\sigma_{\Delta'}$;
- add this bias Δ' to Δ corresponding to the uncertainty on the mean with a fixed data set ;
- compute a random field for residuals using the mean $(\Delta + \Delta')$ and the variance $(1-\sigma^2_{\Delta}-\sigma^2_{\Delta'})$ (4)

With this procedure, the global unit variance of the normal score is split in the different categories of uncertainties that could affect the final map (residual uncertainty, uncertainty on the mean, uncertainty on data with bias).

The graphical illustration of the corresponding multiple realizations loop is depicted on Figure 3.

4 Specific aspects of beta distributions

The beta model distribution (second type) is often used to describe volume ratios as net to gross, but also potentially effective porosities and effective irreducible water saturations. This model is interesting for the following reasons:

- It has bounding values *(min, max)* ;
- the Gaussian model can be seen as a particular case of the beta model
- depending of the shape parameters *(p, q)*, a large variety of behaviour can be described (see Figure 4)

The corresponding distribution law is given by:

$$f(y) = \frac{1}{B(p,q)} . \frac{y^{p-1}}{(1+y)^{p+q}} \qquad \text{with} \qquad B(p,q) = \frac{\Gamma(p).\Gamma(q)}{\Gamma(p+q)} \tag{5}$$

$$y = \frac{x - \min}{\max - \min}$$

In this model, Γ is the gamma function; p and q the shape parameters; they can be derived from the mean *m* and variance σ^2, assuming that bounding values *(min, max)* are fixed and known:

$$p = (m - \min).\left[\frac{(\max - m).(m - \min)}{\sigma^2} - 1\right] \tag{6}$$

$$q = (\max - m).\left[\frac{(\max - m).(m - \min)}{\sigma^2} - 1\right]$$

This relationship allows us to estimate a first guess for fitting the beta law with experimental value of mean and variance. Moreover, in a non-stationary case, a local updating of the shape parameters can be performed from the local mean and variance.

5 Practical case study

All the previous concepts have been used to simulate petrophysical attributes (net to gross, porosity, irreducible water saturation, log Kh, and Kv/Kh ratio) in a carbonate platform reservoir sampled in the hydrocarbon pool with 19 wells.

In this reservoir, five environments of deposition have been distinguished; moreover, the statistics of the net to gross values are different in each of the 15 layers of the model. The corresponding number of fit to achieve is large (450) and cannot be done manually.

The fit of a non-stationary beta law model is performed in two steps:

- computing a vertical trend of mean and standard deviation for each facies,
- derive the shape parameters p and q from local mean and standard deviation using formula (6).

An example of a geostatistical simulation based on this non-stationary model is depicted on Figure 5 (net to gross visualization regarding facies map) and Figure 6 (comparison of statistics between simulation and data)

Hence, the relative uncertainty on the mean values are estimated from the declustering formula (2) using a variogram of 500 meters. This relative uncertainty depends on the facies (some facies are more frequently observed than other) and varies from one variable to another (some variables are less sampled that other, for instance effective water saturation is less sampled than net to gross); for net to gross $r = \sigma_m/\sigma_t$ is between 0.07 and 0.17 (from offshore to upper shoreface facies), for water saturation $r = \sigma_m/\sigma_t$ is between 0.08 and 0.22 (from offshore to upper shoreface facies).

Alternative data set have been produced from a log data uncertainty study. Unfortunately, the systematic bias which exists from possible alternative interpretations have not been correctly represented; to illustrate however the procedure, another example with alternative data set have been generated.

The complete simulation loop process (remember Figure 3) has been used to define uncertainty on volumetrics; it has been compared to the standard case (multiple realizations without uncertainty on data and without uncertainty on means) and to an intermediate case (multiple realizations without uncertainty on data and with uncertainty on the mean).

The volumetrics results are provided on Figure 7. It tends to illustrate that the uncertainty on data are the key uncertainty for this practicle case. This is a consequence of the systematic bias affecting the data; on this mature field, the hydrocarbon pool is controlled by wells and, as a consequence, the uncertainty on mean and the uncertainty on residuals have a small impact on volumetrics. This conclusion is not obvious for reserves and production profiles which are more dependent of local heterogeneities.

6 Conclusions

Uncertainty quantification for hydrocarbon in place evaluation is a frequent exercise in oil industry, this paper has illustrated the possibility of using a geostatistical workflow to achieve this goal. This is not the standard multiple realisation loop frequently observed in commercial software; the suggested procedure involves uncertainty on the distribution model itself, coupled with an uncertainty on conditioning data. Through the case study, it has been shown that this data and mean uncertainties aspect can be a key issue.

It may be argued that the well known technique of experimental design can be used as an alternative approach to treat uncertainty on data and mean. However, multiple realizations are needed to assess local uncertainties on hydrocarbon location; more over, hydrocarbon in place evaluation is not a CPU intensive computation; for all these reasons, it seems more appropriate to use Monte Carlo simulation to explore exhaustively the uncertainty domain instead of focusing on a limited number of cases with experimental design.

Acknowledgements

The author would like to acknowledge J. Gomez-Hernandez (University of Valencia), R. Froidevaux (FSS International), and A. Shtuka (Shtuka Consulting) for stimulating discussions and implementation aspects.

References

Goovaerts, P., *Geostatistics for Natural Resources Evaluation*, Oxford University Press, 1997.

Journel A.G.*, Fundamentals of geostatistics in five lessons*, p.39, Short course in geology: vol.8, American Geophysical Union, Washington DC, 1989

Massonnat G., Biver P., Poujol L., 1998, Key sources of uncertainty from reservoir internal architecture when evaluating probabilistic distribution of OOIP, SPE paper no 49279, *Proceedings of the SPE fall* meeting 1998.

Norris R., Massonnat G., Alabert F., 1993, Early quantification of uncertainty in the estimation of oil in place in a turbidite reservoir, SPE paper no 26490, *Proceedings of the SPE fall meeting 1993*, pp. 763-768.

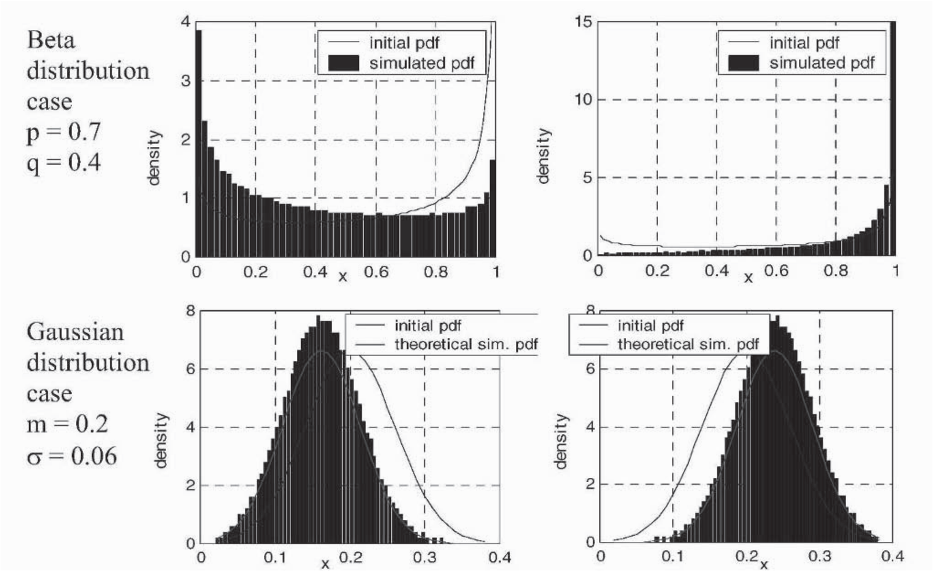
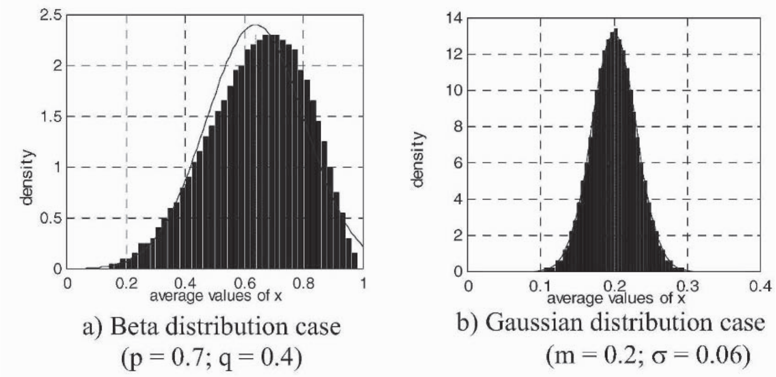


Figure 1: histogram of local values for extreme realisations according to the mean.



a) Beta distribution case (p = 0.7; q = 0.4)

b) Gaussian distribution case (m = 0.2; σ = 0.06)

Figure 2: histograms of the means for the ensemble of realizations.

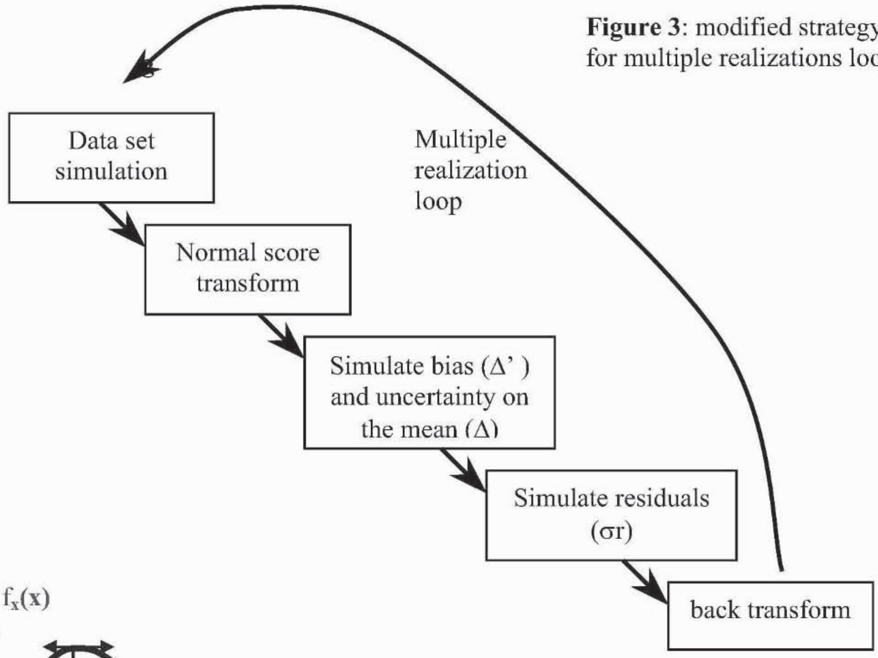


Figure 3: modified strategy for multiple realizations loop.

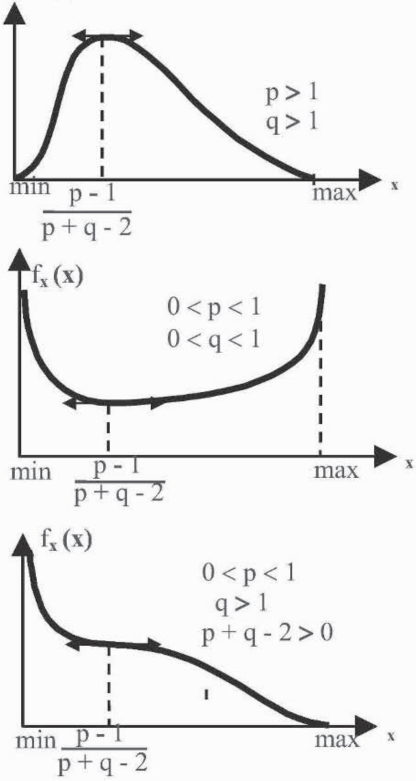


Figure 4: multiple shapes of distributions obtained with various p and q parameters of a beta law model.

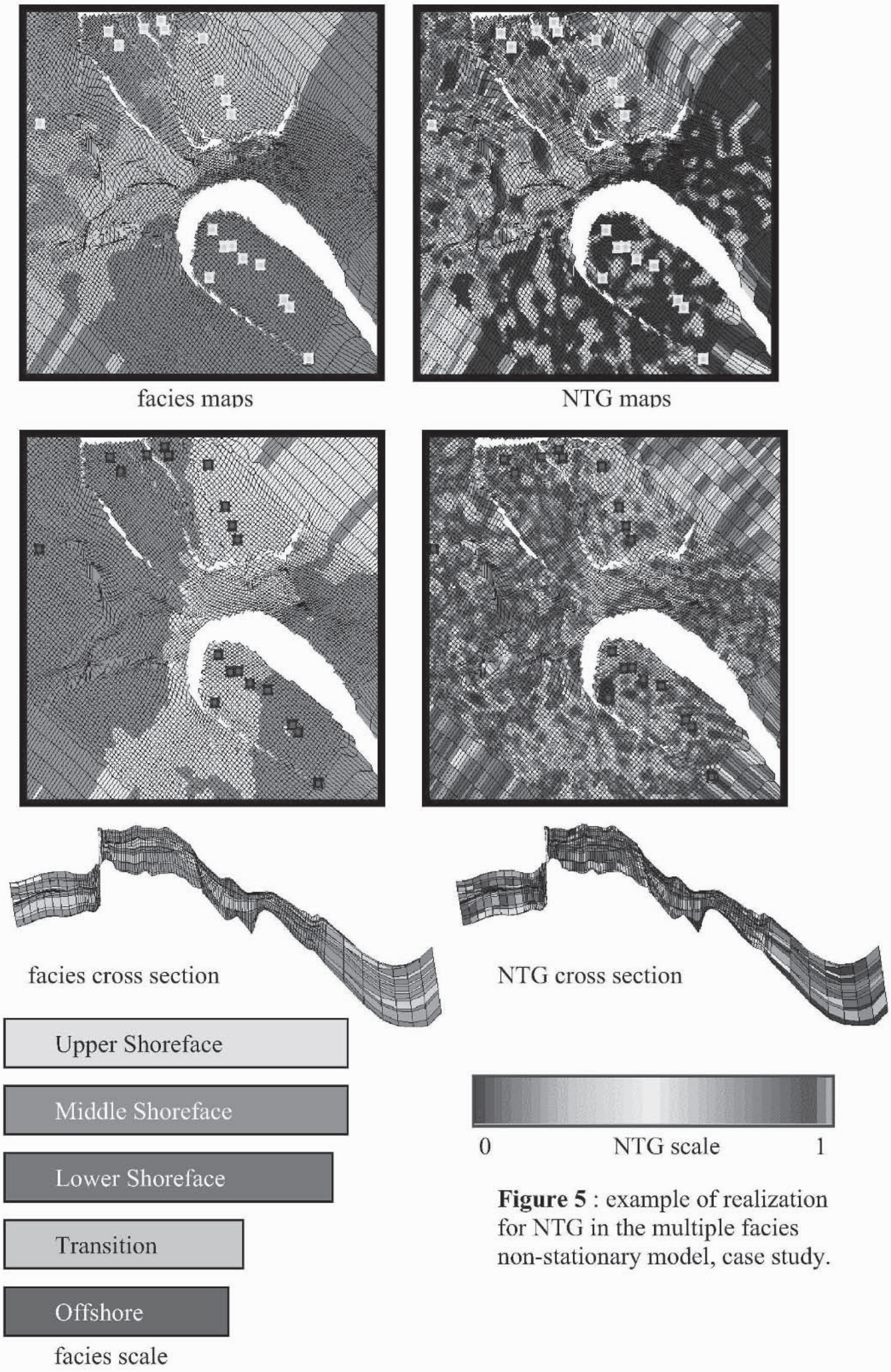


Figure 5 : example of realization for NTG in the multiple facies non-stationary model, case study.

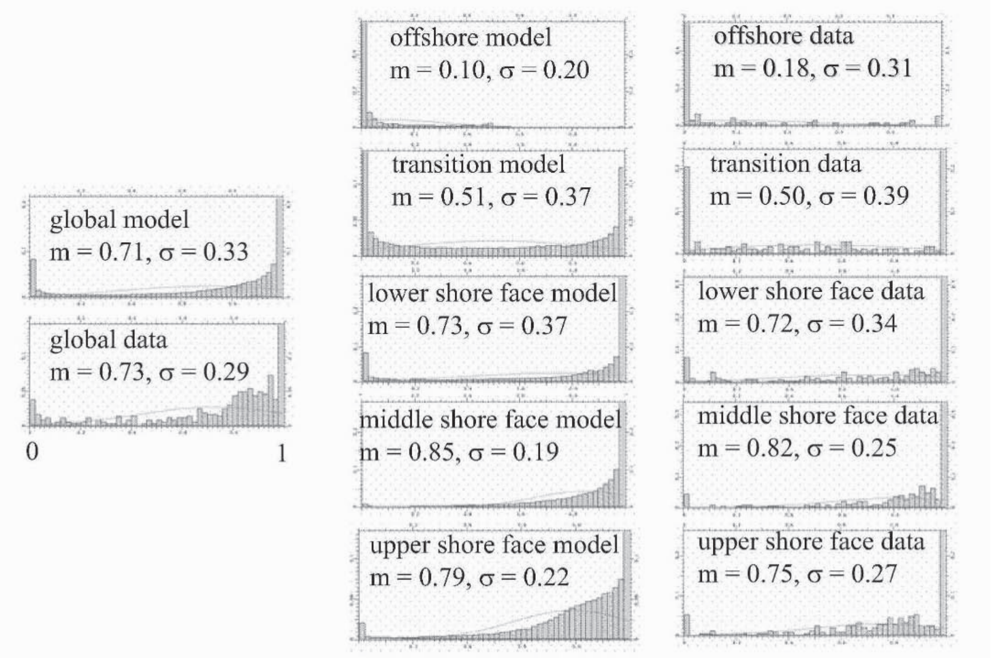


Figure 6: realization statistics versus data statistics, case study, quality control of the model.

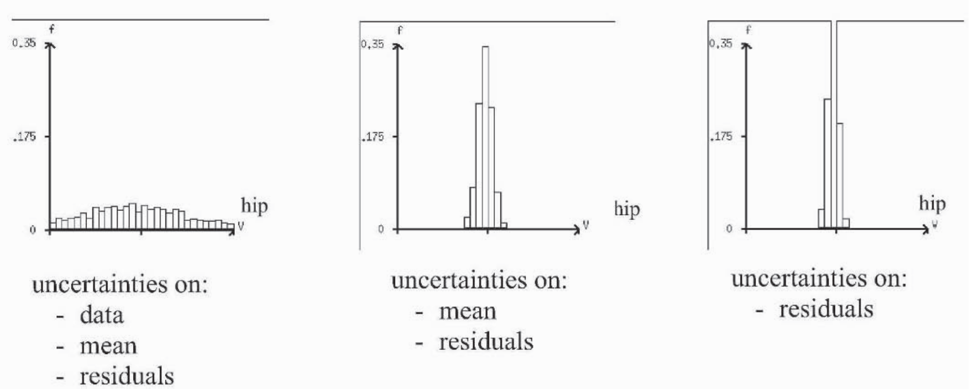


Figure 7: hydrocarbon in place distributions, comparison of runs with different uncertainty sources.

HIGHER ORDER MODELS USING ENTROPY, MARKOV RANDOM FIELDS AND SEQUENTIAL SIMULATION

COLIN DALY
Roxar Limited. Pinnacle House, 17-25 Hartfield Rd., Wimbledon
London SW19 3SE, U.K.

Abstract. Approaches to modelling using higher order statistics from statistical mechanics and image analysis are considered. The mathematics behind a very general model is briefly reviewed including a short look at entropy. Simulated annealing is viewed as an approximation to this model. Sequential simulation is briefly introduced as a second class of methods. The unilateral model is considered as being a member of both classes. It is applied to simulations using learnt conditional distributions.

1 Introduction

For certain problems the entity to be predicted is a non linear function of some poorly known control variable. An example is prediction of flow behaviour in an oil reservoir which involves solving differential equations which are highly non-linear with respect to the spatial distribution of permeability. When the control variable may be modelled stochastically then the accepted practice is to generate several realizations and make predictions by using the distribution of results found by applying the function to each. Assuming that the nonlinear function adequately captures the physics, then our attention turns to seeing if the stochastic modelling adequately captures the geology and to inquire which parts of the geology are relevant for the entity to be predicted.

Two principle methods have been used for simulation, methods based on variograms and object models (although see Tjelmeland and Besag, 1998, for a Markov Random Field approach). The former has tended to focus on simplicity and ease of well conditioning, the latter on geological realism and connectivity. Recently there has been an increase in interest in methods using higher order statistics. There has been some controversy about whether a random function necessarily exists which satisfies these higher order statistics. This paper briefly reviews relevant results from statistical mechanics and image analysis, stating existence results and considering a rigorous general model capable of handling complex higher order statistics. A popular method for using higher order statistics is based on a sequential simulation algorithm (e.g. Strebelle, 2002). This is briefly considered in section 3, which, it should be noted is largely independent of section 2 apart from questions of existence. A unilateral model is a simple special case which is both a Markov random field and yet may be simulated sequentially. For a seemingly non symmetric model it gives surprisingly good results.

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 215-224.

2 Review of results from Statistical Mechanics

2.1 THREE VIEWPOINTS

Suppose we are interested in building stochastic models of a discrete variable (e.g. facies for a petroleum reservoir model) satisfying some vector of statistics η (with given explicit values of η_I - possibly calculated from a training image). Therefore the result of a calculation on a realisation of the model should match these statistics in some sense. The types of calculations (called the *interactions)* that may be used are now defined. They are very wide in scope. Let Λ the grid on which we want to simulate. First we allow *potentials* φ acting on finite subsets of Λ by

(1) $\varphi(X) = \varphi(X + a)$ so that the potential is invariant under translation

(2) $\varphi(\phi) = 0$ where ϕ is the empty set

(3) $\varphi(X) = 0$ for all X such that $diam(X) > R$

The third point says that there is a range R such that the potential acting on a subset bigger than R takes the value zero. The sets for which φ takes non-zero values are called cliques. This restriction can be weakened and is weakened in many of the references given in this section. Then *an interaction* for φ is defined as

$$\Psi = \sum_{X \subset \Lambda} \varphi(X) \qquad (1)$$

The *statistics* that we calculate and compare to the given statistics are just a normalised version of this calculated on a realisation *I(x)*

$$\eta = \frac{1}{|\Lambda|} \sum_{X \subset \Lambda} \varphi(X) \qquad (2)$$

A couple of examples of interactions are

a) Consider Θ to be the set of pairs of points x_1 and x_2 separated by a vector h. In a two facies case, let I(x) be the indicator function for one of the facies. Define a potential by $\varphi_0(X) = \frac{1}{2}$ if $X \in \Theta$ and $I(x_i) \neq \mathrm{I(x_i + h)}$, $\varphi_0(X) = 0$ for any other type of set X. The associated statistic is then just the variogram at lag h.

$$\eta_0 = \frac{1}{|\Lambda|} \sum_{\substack{x_i, x_i + h \\ Z(x_i) \neq Z(x_i + h)}} \frac{1}{2} = \frac{1}{2|\Lambda|} \sum_{x_i} (I(x_i) - I(x + h))^2$$

b) Suppose *I(x)* is a facies model and a permeability value is assigned for each facies. Let X be a set of radius R and let k be the effective permeability calculated in X. Let the histogram of effective permeability for sets of size X be split into n bins. Define potential $\varphi_\iota(X)=1$ if k falls in bin i, and $\varphi_\iota(X)=0$ otherwise. The set of interactions $\{\Psi_\iota\}$ simply records the observed histogram of effective permeability for the realisation *I*.

Our objective will be to be able to do simulations using interactions like $\Psi_0,\ldots,\Psi_n$ by trying to condition to prescribed values of the statistics (where possible). As we can see, the types of interaction and hence statistics that we can work with are very general. From now on we will assume that we are working with vectors of interactions and when we refer to an interaction we will generally mean a vector $\Psi=(\Psi_0,\ldots,\Psi_n)$. We now look at three formulations of this problem which will turn out to be equivalent.

Microcanonical ensemble: Assume that an interaction Ψ is given. This gives rise to statistics η when calculated on a realisation. Let us think of Ω as being the set of all possible realisations on the grid Λ. So if there are n possible facies, then $|\Omega| = n^{|\Lambda|}$ (Actually, we think of the grid as infinite in size in the microcanonical perspective). Let $\Omega(\eta)$ be the subset of Ω taking the statistics η. So this subset contains all the possible realisations with calculated statistics equal to η. Some values of η give large subsets, while others might give small or even empty subsets. The latter correspond to cases where the interaction is not capable of producing the desired statistics. As a simple example, suppose we tried to match $\gamma(h)=-1$ in the first example above. Since the variogram can only be positive, the set $\Omega(-1)$ must be empty. More generally, use of many components in an interaction could lead to contradictions between the components and so to empty subsets. The ideal simulation method for a given set of statistics η would be to sample uniformly from the set $\Omega(\eta)$. Each realisation from $\Omega(\eta)$ takes the same probability but the probability may change for different outcomes n. As such Ω splits into equivalence classes based on the statistics η.

The Gibbs Distribution: In most cases we need to work on a finite grid. The statistics that we calculate on a finite realisation $\eta_\Lambda(I)$ will therefore have some statistical fluctuations. We should not expect to match our input statistics η exactly. The next best thing would be to match them in expectation. The probability distribution that we will choose to work with, call it p, should satisfy

$$\mathrm{E}_p[\eta_\Lambda(I)] = \eta \qquad (3)$$

Defining $s(q) = -\int q(I)\log q(I)dI$ to be the entropy of a distribution q, we will *choose* the distribution p with maximum entropy, that is, $p = \underset{\mathrm{q}}{\mathrm{argmax}}(s(q))$ subject to the constraint (3). Then p follows a Gibbs distribution (see Jaynes 1957, Zhu et al. 1997)

$$p(I;\lambda,\Psi,\Lambda) = \frac{1}{Z_\Lambda(\lambda)}\exp\left(-\langle\lambda,\eta_\Lambda(I)\rangle\right) \qquad (4)$$

where λ are the Lagrange multipliers and $Z_\Lambda(\lambda)$ is a normalisation to ensure the probabilities add to 1. The angle brackets signify scalar product. The λ are found by satisfying equation (3). This is usually a complicated and iterative process.

Markov Random Field: Suppose that we have a neighbourhood system $\{N_i\}_{i\in\Lambda}$, then *I(x)* is a Markov Random Field (MRF) for the neighbourhood system if

$$P[I_i \mid \hat{I}_i] = P[I_i \mid N_i] \qquad (5)$$

where $\hat{I}_i$ means all points on the grid except point *i*. That is to say, the conditional distribution at point *i* given all other points only actually depends on the values in the neighbourhood of *i*.

2.2 EQUIVALENCE OF THE VIEWPOINTS

These three viewpoints give an equivalent mathematical formulation for the problem of trying to produce realisations matching statistics η. We don't attempt to demonstrate the results here as the proofs are long and fairly technical. However we will try to convey the flavour of some of the concepts involved. The question of existence of a random function for an interaction Ψ is treated rigorously in several texts by showing the existence of a *Gibbs measure* (e.g. Ruelle 1968, Georgii 1988). This Gibbs measure has the property that on any finite subgrid, given the values on the exterior of the subgrid, it reduces to the Gibbs distribution on the subgrid. The equivalence of the first and second perspective is called the equivalence of ensembles. The equivalence of the second and third is the Hammersley-Clifford theorem (see e.g. Moussouris 1974). One statement of existence of the random function is given in terms of a new definition of entropy. This definition appears different to the one previously given which was defined as an integral over all realisations using the probability measure of the random function. It is defined below in terms of the size of the set $\Omega(\eta)$ honouring the statistics. Again we take a lax approach to rigor in the following discussion. If $\Omega_{\Lambda}(\eta)$ is the set of realisations on the finite grid Λ taking the statistics η, we define

$$s(\eta;\Psi) = \lim_{|\Lambda|\to\infty} \frac{\log|\Omega_{\Lambda}(\eta)|}{|\Lambda|} \qquad (6)$$

to be the entropy function for the interaction Ψ. The dependence on both Ψ and η might appear confusing, but by the latter we mean the type of calculation that is performed to the realisation while the former are the resulting statistics. We will usually drop the explicit reference to the interaction and just refer to $s(\eta)$. The existence theorem states that for the interactions that we have defined, this limit exists and takes values greater than $-\infty$ for at least some values of η (Ellis 1985). Moreover s is a concave function of η. The definition of s shows us that the number of possible realisations taking statistic η acts like $|\Omega(\eta)| \sim e^{|\Lambda|s(\eta)}$ for large grids. The two definitions of entropy turn out to be equivalent (e.g. Ellis 1985). In fact the entropy $s(\eta)$ is the maximum value that $s(p)$ attains when comparing over all distributions p that satisfy equation 3 (as $\Lambda \to \infty$)

It is important to note how the numbers of realisations depend on the entropy. Suppose we have two possible statistics η_1 and η_2 for the same interaction (concretely, if we think of our calculation of the variogram at lag h giving us two different numbers on different realisations). Suppose that $s(\eta_1)>s(\eta_2)$. Then the proportion of realisations that take value η_1 is $\frac{e^{|\Lambda|s(\eta_1)}}{e^{|\Lambda|s(\eta_1)}+e^{|\Lambda|s(\eta_2)}} = \frac{1}{1+e^{|\Lambda|(s(\eta_2)-s(\eta_1))}} \to 1$. So the number of possible realisations taking higher entropy values grows exponentially higher than those taking lower entropy values.

However this does not yet account for the role of the interaction which supplies a probability distribution on Ω. To see this we use another result (Ruelle, 1968). Starting from the Gibbs distribution perspective (equation 4), it can be shown that a convex function $\rho(\lambda)$ defined below exists. It is called the density function in statistical mechanics

$$\rho(\lambda) = \lim_{|\Lambda\to\infty|} \frac{\log Z_\Lambda(\lambda)}{|\Lambda|} \qquad (7)$$

We can now calculate the probability of a statistic set $\Omega\,(\eta)$. Notice that the Gibbs distribution assigns the same probability to all realisations taking the same statistics η.

Thus we have $P[\Omega_\Lambda(\eta)] = \frac{|\Omega_\Lambda|}{Z_\Lambda(\lambda)}\exp(-\langle\lambda,\eta\rangle)$ By taking logs and passing to the limit and using the density definition we get *the probability rate function,* r_λ (Wu et al. 1999)

$$r_\lambda(\eta) \overset{def}{=} \lim_{|\Lambda|\to\infty} \frac{\log P[\Omega_\Lambda(\eta)]}{|\Lambda|} = s(\eta) - \langle\lambda,\eta\rangle - \rho(\lambda) \qquad (8)$$

We can see that the probability of taking statistic η is asymptotically $P[\Omega(\eta)] \sim e^{|\Lambda| r_\lambda(\eta)}$ so that for large grids the Gibbs distribution samples uniformly from the set $\Omega\,(\eta_{max})$, where $\eta_{max} = \max_\eta(s(\eta) - \langle\lambda,\eta\rangle - \rho(\lambda))$. We know that the entropy function is concave. If, furthermore, it is strictly concave then this implies that only one set of statistics attains the maximum of the probability rate function (Lanford, 1973 gives a sufficient criteria due to Dobrushin based on having sufficiently low values of λ). In other words, the probability concentrates on one particular set of statistics (there is no *phase transition* in the language of statistical mechanics). It turns out that s and ρ are convex conjugate pairs (Lanford, 1973) and using this it is easy to show that for low values of λ (high 'temperature') the probability concentrates on the class which has the highest entropy consistent with the interaction as measured by the entropy function s.

2.3 TWO CONSEQUENCES AND A SIMPLE EXAMPLE

Example: This is a simple example due to Lanford, 1973 in the context of a digression on sums of independent variables. It was instrumental in the reinvigoration of work on the theory of large deviations. Consider a random function that assigns 0 or 1 with equal probability to each point on the grid independently of the value at other points and a potential function that assigns a value 1 to single points, 0 to anything else. Then the statistic of interest is $\eta = 1/|\Lambda|\sum I(x)$, the mean value of the realisation. By applying the binomial theorem and Stirling's formula and taking limits, it can be shown that the entropy is

$$s(\eta) = \begin{cases} -\eta\log\eta - (1-\eta)\log(1-\eta) - \log 2 & 0 \le \eta \le 1 \\ -\infty & \text{otherwise} \end{cases}$$

This takes its maximum value, 0, at $\eta = 0.5$ as expected. So 'virtually all' realisations (elements of Ω) have a mean value of 0.5. Other mean values are possible (between 0 and 1), but in the case of iid random variables they will almost never occur. Since s takes the value -∞ outside [0,1] we get the (obvious) result that the mean cannot take values outside this interval.

Next we look at two consequences:

1) Calculations made on realisations from the same equivalence class $\Omega(\eta)$ give the same results
2) A quick look at 'Simulated Annealing'.

In the petroleum industry it is quite common to build a reservoir model based on some input parameters and test the results or make predictions based on calculations that were not explicitly controlled by the input set. For example we might build a model based on variogram information and validate the model by comparing upscaled permeability values with the observed histogram of upscaled permeability (found by interpreting well test for example). For our purposes we will call such a calculation an *observable o* and demand that the observable satisfies the same constraints as an interaction (e.g. effective permeability calculation over finite regions is an observable as in the earlier example). We know that the realisations of our model will generally sample from $\Omega(\eta_{max})$ where the statistics maximise the probability rate function. We ask how this set decomposes according to the possible values that the observable *o* might take. Well, for each value of *o* there is a subset of $\Omega(\eta_{max})$, call it $\Omega(\eta_{max},o)$. Consider the entropy function for the extended interaction (η_{max},o). It must attain it's maximum for some value of $o_{\max}$. With no phase transition this value is unique. By the exponential growth argument used earlier, the volume $\Omega(\eta_{max},o_{max})$ is far larger than that for other values of o, so $\Omega(\eta_{max})$ is dominated by $\Omega(\eta_{max},o_{max})$ and $s(\eta_{max}) = s(\eta_{max}, o_{max})$. In other words for any interaction and an arbitrary observable the realisations will produce the same statistics for the observable with very high probability. These *typical* realisations have maximum entropy with respect to the extended interaction. Other *atypical* elements of $\Omega(\eta_{max})$ have lower values of the entropy of the extended interaction ('by chance' they have some extra information about *o*) and are comparatively rare so are unlikely to be sampled. So, if two realisations do not have the same statistics on some *o* then it is unlikely that they are from the same equivalence class.

As mentioned before, the Gibbs distribution, given by (4), offers a rigorous method of sampling from a distribution taking a set of statistics. It does so by first finding a set of weights λ for the interaction satisfying the constraints (3). Hence the statistics are satisfied on average. This can be an involved and computer intensive calculation. We now compare this to an application of simulated annealing that has been made regularly in the geostatistics literature, e.g. Deutsch and Journel, 1998. In this method a new Gibbs distribution is proposed (for each value of T).

$$p(\nu) = \frac{1}{Z} e^{-\frac{1}{T}\|\nu-\eta\|} \qquad (9)$$

Here ||.|| is some distance measurement from ν to η. For a fixed value of T, an MCMC technique like the Metropolis algorithm can sample from this distribution. As T→0 the distribution becomes concentrated on those realisations taking the value η, in other words on the set $\Omega_\Lambda(\eta)$. Simulated annealing reduces T slowly to ensure that the sampling is uniform on $\Omega_\Lambda(\eta)$. For large grids, where the statistical fluctuations on the statistics η are small, we can see that this can be viewed as a direct attempt to sample from the microcanonical ensemble. (Technically, we would have to prove that a limit exists as $\Lambda \to \infty$.This is not done here, but it appears to resemble typical statistical mechanical proofs.) A few comments can be made.

1) This will give essentially the same result as sampling from the Gibbs distribution for large grids
2) It can be viewed as an approximate sampling for large grids from the Gibbs measure whose local conditional distributions are the Gibbs distributions (4). In other words, it makes an approximate sample from a regular, well defined random function.
3) For a particular set of data η, which we consider to be associated to an interaction Ψ, there is no guarantee that $\Omega_\Lambda(\eta)$ is nonempty (we have a theorem which says that for any linearly independent set Ψ, there are *some* statistics α for which $\Omega(\alpha)$ is nonempty). If empty, $s(\eta)$ will take the value $-\infty$ in equation (6) and we say that the statistics are incompatible with Ψ. In this case $\Omega_\Lambda(\alpha)\rightarrow 0$ faster than exponentially. This will manifest itself as an inability to match the input statistics to a reasonable degree of accuracy, even for relatively small images. So if the method is producing realisations matching the statistics, the grid is large enough and the annealing schedule is slow, then $\Omega(\eta)$ is non empty and this method should give reasonable results.

3 Sequential Simulation

3.1 GENERAL METHOD

A more detailed analysis of the algorithms in this section is currently in preprint form. The sequential simulation algorithm as proposed in the geostatistics literature, e.g. Deutsch and Journel, 1998, works by assuming known the conditional distribution of a point, x, given any number of neighbours that have been observed within a fixed *search neighbourhood* n_x of the point. The available neighbours have either been previously simulated or are initial conditioning data. We will call these available neighbours the *parents* of x and label the parents as ∂_x and refer to the conditional distribution as $f_d(x\,|\,\partial_x)$. The subscript d alludes to the fact that we are *driving* the simulation by claiming knowledge of some conditional distributions. These distributions may come from an analytic model such as the Gaussian, or they may be empirical distributions coming from some training data. We do not assume, and it is not generally the case, that the simulated model will reproduce all of these statistics. The sequential method then does a simulation on a finite grid by following a path p through the points on the grid. The result follows the distribution (the resultant model does not have the subscript d however it is labelled by p to indicate dependence on the path)

$$f^p(x_1,\ldots,x_N) = f_d(x_1\,|\,\partial_{x_1})f_d(x_2\,|\,\partial_{x_2})\ldots f_d(x_N\,|\,\partial_{x_N}) = \prod_{i=1}^{N} f_d(x_i\,|\,\partial_{x_i}) \qquad (10)$$

Of course, a restriction on this model is that each parent set ∂_x must be known before simulation of x. It is straightforward to show that the conditional distribution is of the form in equation (11) and so depends on the parents of x_i, $pa(x_i)$, the children of x_i, $ch(x_i)$, and the other parents of the children of x_i. $pa(ch(x_i))$

$$f^p(x_i\,|\,\hat{x}_i) \propto \prod_{\{j:i\in\partial_j \; or \;\; j=i\}} f_d(x_j\,|\,\partial_{x_j}) \qquad (11)$$

This set of dependency points is always contained within the *dependency neighbourhood* of x_i, defined as $d(x_i) = pa(x_i) \cup ch(x_i) \cup pa(ch(x_i))$. Thus it is always possible to write

$$f^p(x_i \mid \hat{x}_i) = f^p(x_i \mid d(x_i)) \qquad (12)$$

So that we have a Markov Random Field. The dependency neighbourhood changes for each x_i, so the result is not a stationary MRF. There are two straightforward ways to reintroduce stationarity; using a raster scan path which is the topic of the next section or using a random path as follows. Let P be a random path, that is, a random variable which 'picks' from the set $\wp$ of permutations of *{1,…,N}* uniformly. We now consider the simulation strategy of first picking a path at random and then doing sequential simulation. The distribution f^{P} is a randomisation of that given in (10)

$$f^{\mathrm{P}}(x_1,\ldots,x_n) = \frac{1}{n!}\sum_{p\in\wp} f^p(x_1,\ldots,x_n) \qquad (13)$$

The distribution is now the same for all internal points on the grid. The model still does not reproduce all the input statistics of the driving distribution.

3.2 THE UNILATERAL MODEL

This is a simple model (Picard, 1980) which has the advantage of being both a sequential algorithm and a readily parametrizable MRF at the same time. It reproduces its input statistics and can be simulated in one pass. This gives the advantage of good results (less 'speckle') and a fast algorithm. However, it does depend on initial conditions as we shall see. Let us consider a 2d example to simplify notation. For this model the parent set of any point $x=(x_1,x_2)$ is chosen to be some subset of $\{(a,b); b < x_2 \text{ or } (b = x_2 \text{ and } a < x_1)\}$. We assume that the same subset is always chosen so that the parent set $\partial_x = \partial_{x+h}$ for any h, a translation on the grid. A sequential simulation is made with these parent sets by starting at the top left and finishing in the bottom right. Equation (10) still holds to define the decomposition of the probability distribution. As before, the conditional distribution depends on the parents of x_i, the children of x_i and the other parents of the children of $x_{i.}$ This time however the dependency is the same for all points and the model may be represented as a stationary MRF (see the example below). Simulation of a unilateral model may be made by choosing some initial values along 'the top' of the grid and then simulating in a raster scan order. The fact that we have had to choose some initial values means that the method should be run for a while before it starts sampling correctly from the conditional distribution. An exact sampling, such as that proposed by Propp and Wilson, 1996, appears to be possible but in practice the 'run in' seems to be very short.

An example shows how we get from a representation of the type given by (10) to the Gibbs equivalent.

Example: Consider a 2 facies model where each point x has three parents whose co-ordinates relative to the point are (-1,-1), (-1,0) and (0,-1). Figure 2 labels these points as well as the children and parents of children of *x*. Applying (11) we get (drop the *d*)

$$f(x\,|\,\hat{x}) \propto f(x\,|\,1,2,4)\,f(6\,|\,2,3,x)\,f(9\,|\,4,x,7)\,f(8\,|\,x,6,9) \qquad (14)$$

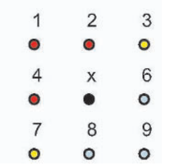


Figure2. Labels for neighbours of x; red = parents; blue = children; yellow = other parents of children

Consider one of the 4 terms on the right, call it *f(y|a,b,c)*, the distribution of *y* given 3 conditioning points. Apart from being a conditional distribution this is arbitrary. We now write a formula for its most general form. Since there are 2 facies, the number of configurations of the conditioning points is 2^3=8. For each configuration we have to specify the probability that *y* takes the value 1 (p(0) = 1-p(1) gives the other value). Let us assume that *f(y|a,b,c)>0* for all combinations of variables (this is not strictly necessary for unilateral process but is for general MRF). Then we can write *f* in the form $f(y\,|\,a,b,c) = \exp(y\Psi(a,b,c))$ where Ψ is an arbitrary function (because y can only take the values 0 or 1 and we only have to concern ourselves with 1 – note this number is not between 0 and 1, but that will be fixed by the normalisation). The most general function of 3 binary variables can be rewritten as

$$\Psi(a,b,c) = \alpha + \beta a + \gamma b + \delta c + \varepsilon ab + \eta ac + \kappa bc + \upsilon abc \qquad (15)$$

by identification of terms, for example $\beta a = \Psi(a,0,0) - \Psi(0,0,0)$. Substituting these into (14) and throwing out terms that do not contain an *x* gives the final MRF result.

$$f(x\,|\,\hat{x}) \propto e^{x\{\alpha+\beta(1+9)+\gamma(2+8)+\delta(4+6)+\varepsilon(12+48+69)+\eta(14+26+89)+\kappa(24+36+78)+\upsilon(124+236+478+689)\}} \qquad (16)$$

This example shows that the unilateral model retains some anisotropy, for example there is no term with (3+7). However, by using larger neighbourhoods any desired terms can be included into the model and we can adequately model complex behaviour. Figure 3 shows a training image and two unilateral simulations of a channel system. The image on the right used a neighbourhood consisting of 60 points. This would appear to necessitate learning 2^{60} configurations for the conditional distribution. While this is true in principle, entropic reasoning tends to suggest that the number of configurations that actually occur is a very small fraction of this. This is not to say that the inference issue is easy. In fact it is the major problem facing techniques trying to use higher order statistics, but in this case we do get a reasonable result (the speckle on the right image is the onset of problems owing to neighbourhood size). Wei and Levoy, 1999 use a unilateral approach with image pyramids to reduce dimensionality.

Conditioning to data introduces some nonstationarity for all sequential simulation methods (conditioning data have no parents – and so have a different conditional distribution to other points). The unilateral method can be made to condition rigorously by using a MCMC on its equivalent MRF formulation. This reduces the efficiency of the unilateral method to that of a typical MRF. The unilateral method offers the possibility of starting with an approximate technique (for example, the algorithm looks to see if any conditioning data are children of the current point being simulated. If so, they are switched and used as parents of the current point). This approximate simulation may be improved by using several iterations of an MCMC algorithm if needs be. Figure 4 shows a conditional simulation using only the approximation technique and no MCMC.

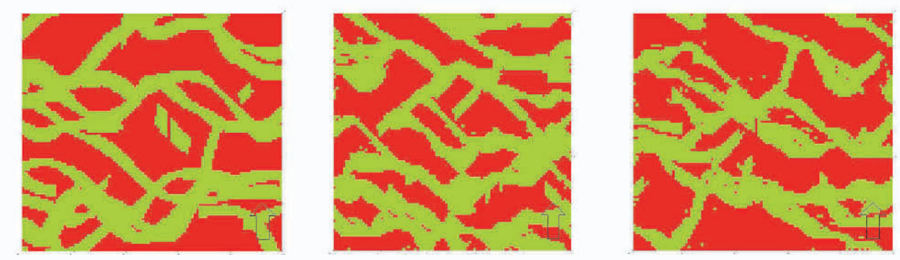


Figure 3. On the left a slice through a boolean channel model. The center and right images are unilateral models with statistics learnt from the left image. The centre model uses a small neighborhood (24 points) while the right one uses 60 points.

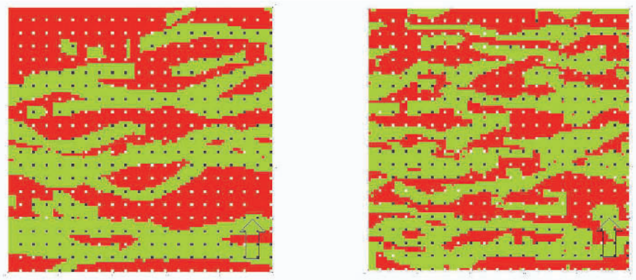


Figure 4. On the left, a slice through a Boolean model and 400 'observed wells' (black and white dots) sampled from the model. The right hand model is a conditional unilateral simulation using the approximation method.

References

Deutsch, C.V. and Journel, A.G., *GSLIB: Geostatistical Software Library and User's Guide*, 2nd Edition, Oxford University Press, 1998.

Ellis, R.S., *Entropy, large deviations and statistical mechanics.* Springer Verlag. New York 1985

Georgii, H.-O., *Gibbs measures and phase transitions.* Walter de Gruyter & Co,. Berlin-New York, 1988

Jaynes, E.T., Information Theory and Statistics, *The Physical Review,* vol.106, no. 4, 1957, pp620-630

Lanford, O.E., Entropy and equilibrium states in classical statistical mechanics *in Statistical Mechanics and Mathematical Problems,*Springer-Verlag, Berlin. Lecture Notes in Physics vol. 20., 1973 pp1-113

Moussouris, J., Gibbs and Markov random systems with Constraints, *Journal of Statistical Physics*, vol. 10, no. 1, 1974, pp11-33.

Pickard, D.K. , Unilateral Markov fields, *Adv. Appl. Probab.*, vol. 12, 1980, pp. 655–671.

Propp, J.G and Wilson D.B., Exact sampling with coupled Markov chains and applications to Statistical Mechanics, *Random Structures and Algorithms*, vol. 9, 1996 pp223-252

Ruelle, D. *Statistical Mechanics: Rigorous results.* W.A. Benjamin Inc. New York-Amsterdam, 1969

Strebelle, S., Conditional Simulation of Complex Geological structures using Multiple-Point Geostatistics, *Math. Geol.,* vol. 34, no. 1, 2002

Tjelmeland H. and Besag J., Markov random fields with higher order interactions, *Scandinavian Journal of Statistics,* vol. 25, 1998, pp 415-433

Wei, L.Y. and Levoy, M., Fast Texture Synthesis using Tree-structured Vector Quantization in *Proceedings of SIG-GRAPH 2000* (July 2000), pp 479–488.

Wu, Y.N, Zhu, S.C and Liu, X., Equivalence of Julesz ensembles and FRAME models, *Int' Journal of Computer Vision*, vol. 38, no. 3, 2000, pp. 245-261,

Zhu, S.C., Wu, Y., Mumford, D., Filters, Random Fields and Maximum Entropy (FRAME), *Int'l journal of Computer Vision,* vol. 27, no. 2, 1998, pp1-20

BEYOND COVARIANCE:
THE ADVENT OF MULTIPLE-POINT GEOSTATISTICS

ANDRE G. JOURNEL
Department of Geological and Environmental Sciences, Stanford CA 94305.

Abstract. In any estimation or simulation endeavor there are two types of information, the conditioning data typically numerical, most often location-specific, and the structural model which relates deterministically or stochastically the conditioning data to the unknown(s). Adding conditioning data is valuable only inasmuch as the structural model that links them to the unknowns is accurate and reflects data redundancy.

Traditional (cross) covariances/variograms, being only 2-point statistics, are limited in the amount of prior structural information they can carry, in addition in 3D they are notoriously difficult to infer and model. In many applications, particularly those related to mapping of categorical variables, facies or rock types distributions, critical structural information can be obtained from training images drawn from prior expertise, outcrops or similar deposits. From such training images complex statistics involving jointly values at multiple locations can be extracted. Using these statistics may be preferable to letting the estimation algorithm impose its arbitrary and likely inappropriate version of the same statistics.

The recently introduced concept of multiple-point (mp) geostatistics allows a fresh methodological look at the general problem of numerical modeling under data conditioning, where the concept of "data" is now open to include structural information much beyond variogram models. Such structural data are often soft and represent a major source of uncertainty, which must and can be appraised through the consideration of alternative training images all consistent with the numerical data available. Training images return more of the modeling responsibility to the geologist, more generally to the physicist who can add interpretation and valuable expertise to the numerical data.

1 From kriging to simulation

Conditional simulation was introduced in the 1970's, interestingly enough not to address uncertainty but as a remedy for the smoothing effect of kriging. It was understood that for many applications reproduction of the patterns of spatial variability, as reflected by the data then modeled through a variogram, was more important than local accuracy. Most of the ensuing developments in the next three decades related to faster and more flexible simulation algorithms, most notably sequential simulation

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 225-233.

algorithms and object-based (Boolean) algorithms and their conditioning to diverse data, both hard and soft.

The practice of simulations quickly revealed the practical limitations of the otherwise extraordinarily congenial Gaussian random function (RF) underlying most simulation algorithms. A single covariance model or matrix was not enough to characterize even the simplest curvilinear structure, any structure that did not display the maximum entropy characteristic of the Gaussian model, a property extraneous to the very concept of a geological structure. Any geostatistician could spot visually and immediately a map generated from Gaussian-related model, any geologist would judge it as relevant only for homogeneously heterogeneous spatial distributions within well-defined homogeneous zones. The concept of indicator RFs extended a little the practicality range of simulation: different categories or classes of a continuous variable could have different variograms. However, in addition to the burden of multiple variograms inference, indicator geostatistics suffered from embarrassing order relation problems, yet did not deliver the flexibility required.

Then came object-based simulations algorithms whereby parametric objects mimicking geological structures were dropped onto the simulation field then moved and morphed iteratively to honor the data. But parametric objects do not offer full flexibility: not all natural shapes can be approximated by simple parametric shapes, and strict conditioning to dense and diverse data was difficult. A new simulation paradigm was needed.

2 From variogram to multiple-point statistics

One reason for the staying power of variogram/covariance-based random function models could be the sense of objectivity one felt at estimating the structural model from actual data. Unfortunately, that sense is more illusion than reality. First, except for data-rich fields, variograms are notoriously difficult to infer then model, to a point that, in petroleum applications for example where hard data are scarce, that task is often not even attempted: variograms are synthesized from expertise, distant outcrops or loosely related ancillary data such as provided by seismic surveys. Second, what really matters for simulation is the random function model adopted, more precisely its multivariate or multiple-point (mp) distribution, not its variogram which is but a two-point statistics; one notable exception is (precisely!) the Gaussian model. Different RF models sharing the same variogram model and honoring the same data values at the same locations could yield drastically different simulated realizations, see Figure 1. The variogram has little structural resolution, cannot distinguish vastly different patterns of heterogeneity. Consequently the variogram is also an incomplete measure of uncertainty; the major source of uncertainty does not lie in fluctuations of various conditional realizations sharing the same variogram model but in the choice of the generating RF model much beyond its variogram, see Caumon and Journel (2004).

But the consideration of mp statistics raises the issue of inference. If variograms are already difficult to infer from actual data, there is no hope to infer even an elementary 3-point statistics, let alone multiple-point statistics. Under stationarity, that inference would require availability of multiple replicates of triplets of data sharing the same geometric configuration, that is the same pair of separation vectors (**h**$_1$, **h**$_2$), as opposed to a variogram requiring replicates of only doublets of data sharing the same single separation vector **h**. One had to overcome the illusion of objectivity provided by direct inference of statistics and accept that multiple statistics could be inferred from training images (TIs), which in many practices was already done for the variogram. Inferring a variogram from a noisy experimental variogram cloud then accepting blindly the higher order statistics implicit to, say, a Gaussian RF model is no more objective than inferring the same high order statistics from a visually explicit training image. A training image can be refuted by an expert geologist; the same cannot be said about a geologically non-significant variogram model.

3 From mp statistics to direct probability inference

Introduction of the concept of mp statistics and accepting that they could be inferred from training images led to the very demise of these statistics. In the first instance, why do we need a variogram? The variogram model is used to build the various conditional probability distributions from which simulated values are drawn. If 3-point, 4-point, mp statistics were available, one could derive from them better conditional probabilities, better in the sense that the resulting simulated values would reflect these higher order statistics in addition to the variogram. But as to infer these high order statistics from a training image why not infer from the same training image and directly the required conditional probabilities? Identifying conditional probabilities to conditional proportions read directly from the training image would shortcut completely the step of moments inference and modelling and that awkward step of kriging which relates the variogram model to the conditional probability. In practice, one would scan the training image for replicates of the multiple-point conditioning data event; these replicates would provide a distribution of the corresponding central training values; that distribution is then taken as the conditional probability. It can be shown that if a training image is considered a representation of an ergodic random function, its training proportions identify exactly the conditional probabilities that would be calculated from the experimental moments lifted from the same training image. If the conditional probabilities are available directly, why take the indirect and painful route of inferring all the relevant high order moments to reconstitute exactly the same probabilities through some kriging?

This remarkable leap of thought was due to Srivastava (Guardiano and Srivastava, 1992). Srivastava's original implementation required, however, to scan the training image repetitively for each new unsampled location, an overwhelming cpu task if large fields with 10^6 to 10^8 nodes are to be simulated. Then came faster desktop computers with larger RAM and the contribution of Strebelle (2000, 2002). Strebelle suggested to scan the training image only once with a specific data template (size and geometry), record in RAM all training data events together with their central training values. The search tree data structure used for that record allows a fast retrieval of all required conditional probabilities. Strebelle and large RAM availability made the new paradigm of Srivastava practical.

Figure 1 gives three very different spatial distributions, yet sharing the same histogram (here a proportion) and, most significantly, approximately the same variogram. Thus any solely variogram-based simulation algorithm would fail to resolve the structural difference between these training images. Figure 2 gives three conditional simulations using Strebelle's snesim code, each drawn from a different training image (as given in Figure 1), but conditioned to the same 30 data values and the same global proportion. Each mp-based conditional simulation reproduces fairly well its training image and, most significantly, the common variogram model even though no such model was ever input into the mp simulation code. The mp simulation algorithm also never used any kriging, although one could argue that the identification of the conditional probability to the training proportion amounts to solve a kriging system with a single (normal) equation, hence the name snesim (single normal equation-based simulation) given by Strebelle to his algorithm.

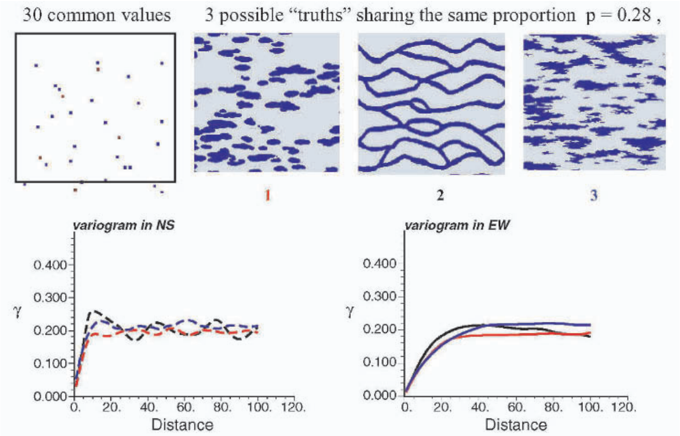


Figure 1. The need to go beyond the variogram. The variogram model cannot resolve the 3 possible "truths"

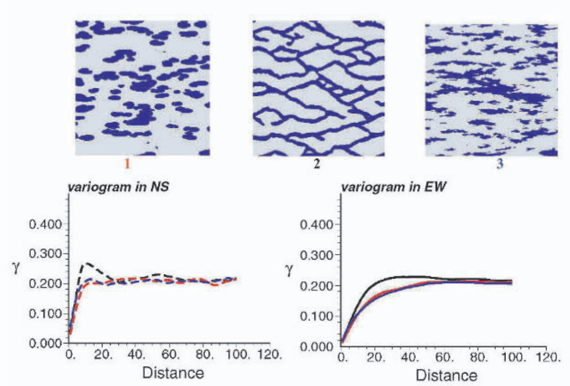


Figure 2. mp simulations using the training images of figure 1. They are conditioned to the same 30 samples and global proportion p=.28

Figure 3 gives a mp simulation generated from a training image which is a realization of a variogram-based sequential Gaussian algorithm. Again no variogram model was ever considered, yet the mp realization generated succeeded to reproduce the training image variogram, thus proving that mp simulation could replace traditional variogram-based algorithms as long as the relevant training image is available. Because no kriging system had to be built and solved, the mp generation of that Gaussian realization was faster cpu-wise than the generation of the training image using the traditional Gaussian sequential simulation algorithm. This remark points to the idea of a "universal" catalog of training images that would include all typical geological structures, a class of which being that of maximum entropy Gaussian-type structures. Once an appropriate training image is retrieved from that catalog, mp simulation with data conditioning could be lightning fast without the encumbrance of variogram modelling and kriging.

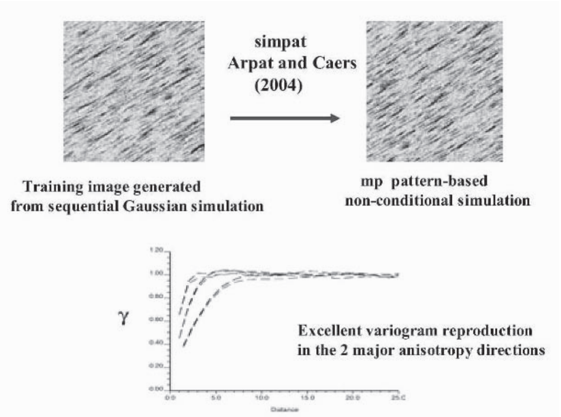


Figure 3. Pattern-based simulation of a continuous variable

4 From point simulation to pattern simulation

Change of paradigm breeds accelerated advances. Since the introduction of the snesim algorithm by Strebelle, within a period of a few years, many other mp simulation algorithms have been developed, some already in a state of beta testing, Arpat and Caers (2004), Zhang et al. (2004). As for lifting from a training image probability distributions of central point values conditioned by a mp data event, why not lift probabilities of whole multiple-point patterns conditioned to the same mp data event? A pattern, that is a mp event, would be drawn from a certain class of training patterns and patched onto the simulated field with due consideration to conditioning data.

Pattern simulation is based on the same concept: infer distributions directly from training images shortcutting all steps of elementary statistics inference and reconstruction of conditional probabilities. Stochastic simulation becomes an exercise of image construction drawing from a set of training puzzle pieces:

1) training patterns are first classified in bins according to some similarity/distance criterion. Each bin is characterized by an average pattern called prototype.
2) define a path (typically random) visiting all unsampled nodes of the field to be simulated.
3) at any location along that path, collect its mp conditioning data event, find the training prototype most similar to that mp data event, and draw a training pattern from that prototype bin.
4) patch that pattern onto the simulation field overriding any non hard data, a hard data being either original data or previously simulated values marked as hard. All values of that pattern becomes conditioning data for calculating distances but only the central part of that pattern is marked as hard data never to be changed.
5) move to the next non hard data location along the simulation path and repeat the pattern simulation procedure until the path is completed. A simulated conditional realization has been generated.

Figure 4 gives an example of such mp pattern simulation. At the top of the Figure is the training image, a binary image of dry soil cracks; 50 hard data are taken from that image. At the bottom of the figure are given 3 conditional simulations using the recently developed mp code filtersim (Zhang et al., 2004). These simulations honor exactly the 50 data at their locations, and reproduce reasonably the training image patterns although with shorter less connected cracks, a generic problem associated with the Nyquist frequency limitation.

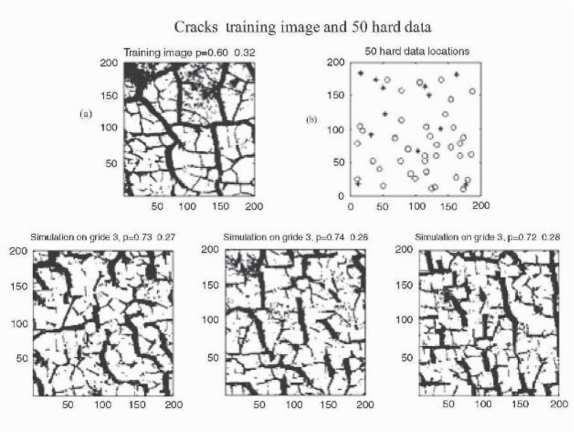


Figure 4. Three pattern-based conditional simulations

5 Issues and challenges

mp geostatistics started by recognizing the limitation of the variogram/covariance tool which are mere two-point statistics. Next the value of prior structural information lifted from training images was recognized. mp higher order statistics much beyond the variogram could now be inferred, but then also and directly the conditional probabilities needed for simulation. Direct lifting of these conditional probabilities voids the need for inference of any other statistics, whether a variogram or else, and also voids the need of reconstructing the same probabilities by indicator kriging or any other algorithm.

One could see a training image as the representation of an ergodic RF, or be bold enough to shed all probabilistic reference and take the training image for what it is truly, a repository of training patterns from which similar looking images anchored to prior data can be built. The puzzle reconstruction draws from a box with many bins, each bin filled with "similar-looking" patterns, the rule being that any pattern used is immediately replaced by a twin in its bin. From a given training image, for a given set of conditioning data, there are many alternative reconstructions possible: this is the within-model uncertainty. Many alternative training images could be considered, which amounts to retaining different puzzle boxes and results in different sets of simulated images: this is the model uncertainty, typically much larger than any within-model uncertainty, Caumon and Journel (2004).

There are definite issues in how to select a training image type then building that training image if it is not already available in a catalog, but this is no different from choosing a variogram model and the RF spatial law implicit to any specific simulation algorithm. There is no escaping from adopting a RF model the very moment one draws a map or contour a line, even if this is done by hand. A training image allows utilizing much richer prior structural information, as could be obtained from experience, outcrops or analog fields. It would be foolish to ignore such valuable information, information that a mere variogram cannot carry.

Then there is the issue of model and data consistency. If the training image is inconsistent with the local conditioning data, and if the mp simulation algorithm freezes the latter, discontinuities will be simulated next to these data. Such discontinuities may not be spotted if there are few conditioning data. Indeed just about any structural model could be anchored to sparse data, the model uncertainty is then overwhelming, as should be expected. Figure 5 gives a large (5×10^5 nodes) 3D pattern-based simulation of a channel reservoir conditioned to a large number of well data: the result is impressive because the training image reflects accurately the heterogeneity patterns of the actual reservoir, in particular the vertical stacking of the channels. Figure 6 repeats that exercise but now using a poorer training image which fails to display the channel vertical stacking: that inconsistency with the well data leads to a much poorer simulated realization: the realization honors the well data but with discontinuities.

A maximum entropy structural model a la Gaussian or a la high nugget effect would be much more permissive as for data inconsistency thanks, precisely, to its high entropy/disorder which crowds data discontinuities. Such tolerance is, however, dangerous because it masks problems.

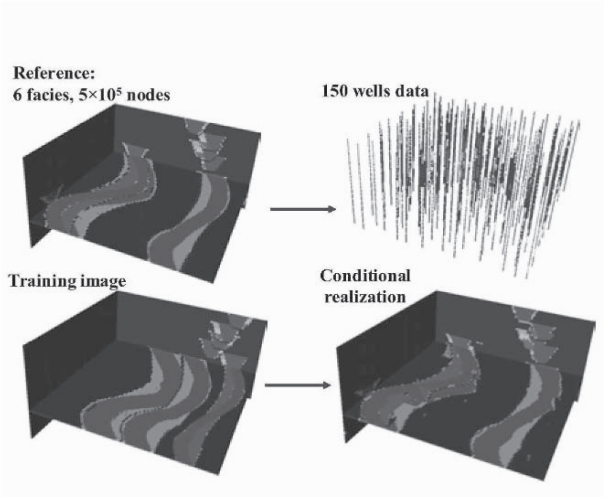


Figure 5. 3D pattern-based simulation with consistent data

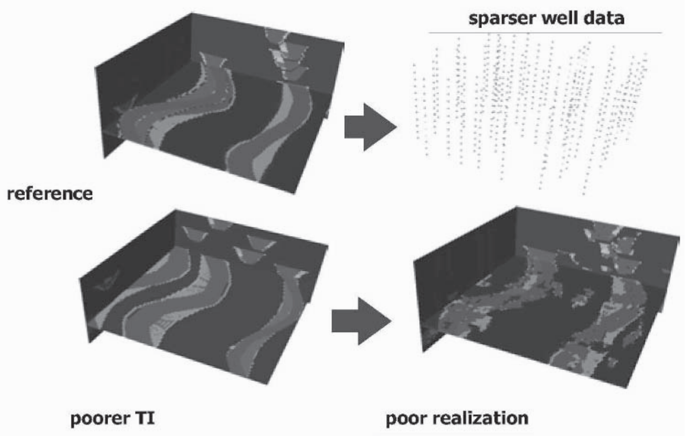


Figure 6. Data inconsistent with the training image

6 Conclusions

The inception of the multiple point concept has brought major changes into the way one sees spatial modelling. With the remarkable processing power now available on desktop computers, there is no more reason to ignore critical prior structural information brought by human expertise under the argument that it is subjective or “messy”, not delivered concisely through a few statistics. Just like the original development of geostatistics in the 1960’s was made possible by the availability of digital computing (variogram calculation and solving kriging systems), now is the time to graduate into massive processing of expert structural information which will tie the discipline of geostatistics increasingly more to image and computer sciences.

A model should never take precedence on data, or analytical convenience on completeness of the model. Real phenomena can rarely be summarised into a few simple statistics (e.g., a histogram and a variogram), any complexity that matters must be faced outright calling on all sources of information available to model it. It is counterproductive to ignore expert interpretative data because they are soft. A variogram model delivers little, yet its inference from typically too few data makes it as soft as a well documented training image that carries much more valuable information. As any other data, training images are uncertain and can be considered as random variables; a distribution of alternative training images might be considered, thus adding an essential component of uncertainty to the final prediction. The implicit high entropy training image built in the traditional mapping algorithms is no less uncertain than the visually explicit training image provided by an expert geologist.

References

Arpat, B. and Caers, J., *A multiple-scale pattern-based approach to sequential simulation*, In Proc. of the 2004 Banff Geostatistics Congress, Kluwer publ., (this volume), 2004

Caumon. G. and Journel, A.G., *Early uncertainty assessment: Application to a hydrocarbon reservoir development*, in ibid, 2004

Zhang, T., Switzer, P. and A. G. Journel, *Sequential conditional simulation by identification of training patterns*, in ibid, 2004

Guardiano, F. and Srivastava R. M., *Multivariate geostatistics: Beyond bivariate moments*. In Geostat Troia 1992, ed. Soares, Kluwer publisher, 1992

Strebelle, S., *Sequential simulation drawing structures from training images*, unpublished PhD Thesis, Stanford University, 2000

Strebelle, S., Conditional *simulation of complex geological structures using multiple-point statistics*, Mathematical Geology, Vol. 34, no. 1, 2002, p. 1-22.,

This page intentionally blank

NON-STATIONARY MULTIPLE-POINT GEOSTATISTICAL MODELS

SEBASTIEN STREBELLE(1) and TUANFENG ZHANG(2)
(1) ChevronTexaco Energy Technology Company,
6001 Bollinger Canyon Road, San Ramon, CA 94583, USA
(2) Department of Geological and Environmental Sciences
Stanford University, Stanford, CA 94305, USA

Abstract. During the last few years, the use of multiple-point statistics simulation to model depositional facies has become increasingly popular in the oil industry. In contrast to conventional variogram-based techniques such as sequential indicator simulation, multiple-point geostatistics enables the generation of facies models that capture key depositional elements (e.g. curvilinear channels) characterized by unique and predictable shapes. In addition, multiple-point geostatistics is more intuitive because the complex mathematical expression of the variogram is replaced with an explicit three-dimensional training image that depicts the geometrical characteristics of the expected facies.

In multiple-point geostatistics, the stationarity assumption that underlies the inference of a variogram model from sparse sample data is extended to infer facies joint-correlation statistics from the training image. A consequence of this assumption is that patterns extracted from the training image can be reproduced in any region of the reservoir model where the training image is thought to be representative of the geological heterogeneity. Yet actual reservoirs are generally non-stationary: topographic constraints, sea-level cycles, or changes of sedimentation sources lead to spatial variations of facies deposition directions and facies geobody dimensions. Three-dimensional fields of location-dependent facies azimuth/dimensions representing those spatial variations are commonly estimated from well log and seismic data, or from geological interpretations based on analogs. This paper proposes a modification of the multiple-point statistics simulation program **snesim** to account for such non-stationary information.

In the original **snesim**, prior to the simulation, the multiple-point-point statistics inferred from the training image are stored in a dynamic data structure called a search tree. In the presence of a locally-varying azimuth field, the range of possible azimuths over the study field is first discretized into a small number of classes. Then, the training image is successively rotated by the average value of each azimuth class and a search tree is built for each resulting rotated training image. During the simulation, at each unsampled node, multiple-point statistics are retrieved from the search tree built for the class in which the local azimuth falls, enabling the local reproduction of patterns similar to those of the corresponding rotated training image. A similar process is proposed to account for a field of location-dependent facies geobody dimensions. The new modified **snesim** program is applied to the simulation of a fluvial reservoir with locally-variable channel orientations and widths.

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 235-244.

1 Introduction

Multiple-point geostatistics has emerged recently as a practical approach to characterize and model facies at reservoir scale (Strebelle *et al*, 2002). The first step of this approach is the construction of a three-dimensional training image describing the facies thought to be present in the study area. The training image captures the geometrical characteristics of each facies, as well as the complex spatial relationships among multiple facies. The training image is a purely-conceptual geological model; it contains no absolute location information and in particular, is not conditioned to any actual field data. In reservoir modeling applications, non-conditional object-based modeling techniques appear to be well-suited to create such three-dimensional conceptual models. The second step of this approach consists of inferring from the training image statistics on the joint-correlation of facies at multiple locations, and using these statistics to reproduce patterns similar to those of the training image while honoring hard and soft conditioning data.

The theoretical framework of multiple-point geostatistics was developed as early as 1989 by Journel and Alabert and was revisited by Guardiano and Srivastava in 1993. The first practical implementation was proposed by Strebelle (2000), who introduced a dynamic data structure called a search tree, to efficiently store and retrieve all multiple-point statistics inferred from the training image. During the last few years, multiple-point geostatistics has been shown to overcome the major limitations of traditional facies modeling technologies:

- Multiple-point statistics (MPS) simulation enables improved modeling of curvilinear and large-scale continuous facies patterns, such as sinuous channels, relative to variogram-based techniques (Strebelle *et al*, 2002). In addition, the training image is much easier to analyze/discuss than a variogram model.
- In contrast to object-based modeling techniques (Holden *et al*, 1996; Viseur, 1997; Lia *et al*, 1998), MPS simulation is a very flexible data integration tool. In particular, MPS models honor all conditioning well data, i.e. reproduce at all well data locations the facies connectivity/geometry observed in the training image, with no limitation on the number of wells (Strebelle and Journel, 2001).

One important assumption underlying the inference of multiple-point statistics from the training image and their reproduction in the MPS model is the stationarity of the field under study: facies relative proportions, geometries, and associations are expected to be reasonably homogeneous over the field. Yet, most actual reservoirs are not stationary. Local topographic constraints such as the presence of a salt dome, seal level cycles, or changes of sedimentation sources, lead to significant spatial variations of facies deposition directions and facies geobody dimensions.

In this paper, we first review the implications of the stationarity assumption in multiple-point geostatistics. Then we propose modifying the MPS simulation program **snesim** (Strebelle, 2000) to reproduce pre-defined non-stationary information such as locally-varying facies azimuth and/or facies geobody dimension data.

2 Stationarity

Geostatistics relies on the concept of Random Function. The Random Function represents the statistical model of spatial variability of some property over some study field. In traditional geostatistics, the Random Function model is generally limited to some one-point and two-point statistics moments, namely a cumulative distribution function and a variogram model. In multiple-point geostatistics, the Random Function model consists of the multiple-point facies joint-correlation moments that can be inferred from the training image. The inference of statistics representing the Random Function model requires some repetitive sampling. For example, a porosity cumulative probability distribution is typically inferred from the histogram of porosity data collected from all well logs available over the study field. However, when pooling sample data together into a single histogram, the modeler makes an assumption of stationarity: all porosity sample values are assumed to originate from the same unique population, regardless of their location in the reservoir. Another stationarity decision is commonly taken whenever a variogram is computed by pooling information at similar lag distances together into a single scatter plot.

In multiple-point geostatistics, the stationarity assumption carries over to higher order statistics: multiple-point statistics moments are inferred from training patterns present in the training patterns regardless of the location of these patterns in the training image. As a consequence, non-stationary features of the training image cannot be preserved in MPS models. Figure 1 shows a clearly non-stationary training image wherein ellipses are South West-North East-oriented in the left half of the image, and North West-South East-oriented in the right half. The resulting model generated by the MPS simulation program **snesim** displays a mix of ellipses oriented in both directions over the whole field.

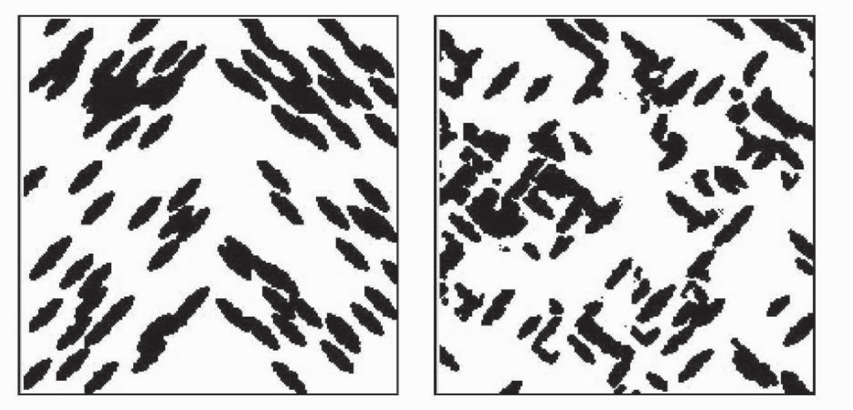


Figure 1. Non-stationary training image (left), and resulting MPS model (right). The specific locations of the South West-North East and North West-South East-oriented ellipses in the training image are not preserved in the MPS model.

The non-stationary features of the training image are not captured in MPS models. Therefore, we propose using a stationary training image and applying rotation and affinity transforms to the training image to reproduce non-stationary features to MPS models. Prior to that, the implementation of the original MPS simulation program **snesim** is briefly recalled.

3 Multiple-point statistics simulation implementation

The MPS simulation program **snesim** proposed by Strebelle (2000) is a pixel-based direct sequential simulation algorithm: all simulation grid nodes are visited only once along a random path and simulated node values become conditioning data for cells visited later in the sequence. Let S be the categorical variable (depositional facies) to be simulated, and s_k, k=1…K, the K different states (facies types) that the variable S can take. At each unsampled node $\mathbf{u}$, d_n denotes the data event consisting of the n conditioning data $S(\mathbf{u_1})=s(\mathbf{u_1})\ldots\ S(\mathbf{u_n})=s(\mathbf{u_n})$, closest to $\mathbf{u}$. The conditional probability distribution function (cpdf) at $\mathbf{u}$ is inferred by scanning the training image to find all training replicates of d_n (same geometric configuration and same data values as d_n), and identifying the conditional facies probabilities as the facies proportions obtained from the central values of the training d_n –replicates.

Instead of repeatedly scanning the whole training image at each unsampled node to search for training replicates of the local conditioning data event, Strebelle (2000) proposed storing ahead of time all conditional facies probabilities that can be inferred from the training image in a dynamic data structure called a search tree. More precisely, given a conditioning data search window W, which may be a search ellipsoid defined using GSLIB conventions (Deutsch and Journel, 1998), τ_N denotes the data template (geometric configuration) constituted by the N vectors $\{\mathbf{h}_\alpha, \alpha=1\ldots N\}$ corresponding to the N relative grid node locations included within W. Prior to the simulation, the training image is scanned with τ_N, and the numbers of occurrences of all training data events associated with τ_N are stored in the search tree. During the simulation, at each unsampled node $\mathbf{u}$, τ_N is used to identify the conditioning data located in the search neighborhood W centered on $\mathbf{u}$. d_n denoting the data event consisting of the n conditioning data found in W (original sample data or previously simulated values, $n \leq N$), the local probability distribution conditioned to d_n is retrieved directly from the above search tree; the training image need not be scanned anew.

Theoretically, a large data template τ_N should be used to capture the large-scale features of the training image. However, such large template would increase dramatically the memory used to build the search tree and the cpu-time needed to retrieve conditional probabilities from it. One practical solution to capture large-scale structures while keeping the size of the data template τ_N reasonably small ($N \leq 100$) is to use a multiple grid simulation approach (Strebelle, 2000). In **snesim**, this approach consists of simulating a series of G increasingly-finer grids, the g-th ($1 \leq g \leq G$) grid comprising each $2^{\mathrm{G-g}}$-th node of the final (finest) simulation grid. After the data template $\tau_N=\{\mathbf{h}_\alpha, \alpha=1\ldots N\}$ has been defined on the finest grid, its components $\mathbf{h}_\alpha$ are rescaled proportionally to the node spacing within the grid being simulated. Thus the rescaled data template $\tau_N^{g}=\{\mathbf{h}_\alpha^{g}=2^{\mathrm{G-g}}.\mathbf{h}_\alpha\ , \alpha=1\ldots N\}$ is used to build the search tree and search for conditioning data when simulating the g-th grid.

In the next two sections, we show how rotation and affinity transformations can be applied to the data template τ_N prior to building the search tree, to integrate location-dependent azimuth and geobody size information into MPS models.

4 Integration of azimuth data

In this section, we first study the simple case in which the main direction of continuity of the facies geobodies is assumed to be constant over the field under study, but possibly different from the main direction of continuity of the training facies. Then we extend this technique to handle 2D or 3D fields of location-dependent azimuths. Only azimuths defined in the *xy*-plane are considered in this section because, in practice, dip is typically taken into account by the layering of the stratigraphic grid in which the facies model is built.

4.1 CONSTANT AZIMUTH

Consider the case in which the facies geobodies in the MPS model should have a constant principal direction of continuity, yet possibly different from that of the training image. Let θ be the difference in degrees counter-clockwise between those two directions.

Given a training image and a data template $\tau_N=\{\mathbf{h}_\alpha, \alpha=1\ldots N\}$, Zhang (2002) proposed modifying the **snesim** algorithm as follows. First the search tree is built from the training image using τ_N. Then, a new data template $\tau_N(\theta)$ is created from τ_N by the following method:

1. Rotate by θ each single component $\mathbf{h}_\alpha$ of τ_N.
   In 2D, the coordinates $(x_\alpha(\theta), y_\alpha(\theta))$ of the rotated component $\mathbf{h}_\alpha(\theta)$ are computed from the coordinates (x_α, y_α) of the original component $\mathbf{h}_\alpha$ as:
   $x_\alpha(\theta)= x_\alpha \cos\theta + y_\alpha \sin\theta$
   $y_\alpha(\theta)= - x_\alpha \sin\theta + y_\alpha \cos\theta$
2. Relocate the rotated components $\mathbf{h}_\alpha(\theta)$ to the nearest nodes of the simulation grid currently simulated.

During the simulation, at each unsampled node, the rotated data template $\tau_N(\theta)$ is used to search for nearby conditioning data, and the corresponding conditional probability distribution function (cpdf) is retrieved from the search tree, which was built using the original data template τ_N.

However, as described in the previous section, **snesim** uses a multiple-grid simulation approach that consists of simulating a series of increasingly-finer grids. Thus, at the early stage of the simulation, the rotated components $\mathbf{h}_\alpha(\theta)$ are relocated to the closest nodes of some coarse grids, entailing drastic approximations regarding the actual locations of the conditioning data. Such approximations lead to the inaccurate estimation of facies probability distributions and the poor reproduction of training patterns. However, because the simulation grids used in **snesim** are regular Cartesian grids and the distance between nodes is the same along both x and y-directions, the components $\mathbf{h}_\alpha(\theta)$ of the rotated data template match exactly existing grid nodes for θ=0, 90, 180, or 270 degrees. This is a property that we will use in the next sub-section.

An alternative approach consists of keeping the original data template τ_N to search for conditioning data, but rotating that data template to build the search tree prior to the simulation. In this case, the rotated data template is built in a slightly different way than in Zhang's method:

1. Rotate by $-\theta$ each single component $\mathbf{h}_\alpha$ of the original data template τ_N.
   In 2D, the coordinates $(x_\alpha(-\theta),\ y_\alpha(-\theta))$ of the rotated component $\mathbf{h}_\alpha(-\theta)$ are computed from the coordinates (x_α, y_α) of the original component $\mathbf{h}_\alpha$ as:
   $x_\alpha(-\theta)= x_\alpha \cos\theta - y_\alpha \sin\theta$ and : $y_\alpha(-\theta)= x_\alpha \sin\theta + y_\alpha \cos\theta$
2. Relocate the rotated components $\mathbf{h}_\alpha(-\theta)$ to the nearest nodes of the training image grid. When using **snesim**, the training image is assumed to have the same node spacing as the (finest) simulation grid.

The resulting rotated data template $\tau_N(-\theta)$ is used to build the search tree from the training image. The exact same result can be obtained by rotating the training image by θ, then building the search tree from that rotated training image using the original data template τ_N. During the simulation, at each unsampled node, τ_N is used to search for nearby conditioning data, and the corresponding cpdf is retrieved from the above search tree.

The most critical advantage of that new technique over Zhang's original method is that the relocation of the rotated components $\mathbf{h}_\alpha(-\theta)$ to the training image grid entails only minor approximations of the actual locations of the conditioning data, thus resulting in a reasonably good reproduction of the training patterns. As an application, this modified **snesim** program was used to model a horizontal 2D section of a fluvial reservoir. The training image depicts the prior conceptual geometry of the sinuous sand channels expected to be present in the subsurface (Figure 2). The size of that image is 250*250=62,500 pixels, and the channel proportion is 27.7%. A non-conditional simulated realization was generated using the same direction of continuity as that of the training image (Figure 2), then two additional models were created using different arbitrary main directions of continuity: 20 and 50 degrees (Figure 3). All models reproduce equally well the patterns displayed in the training image.

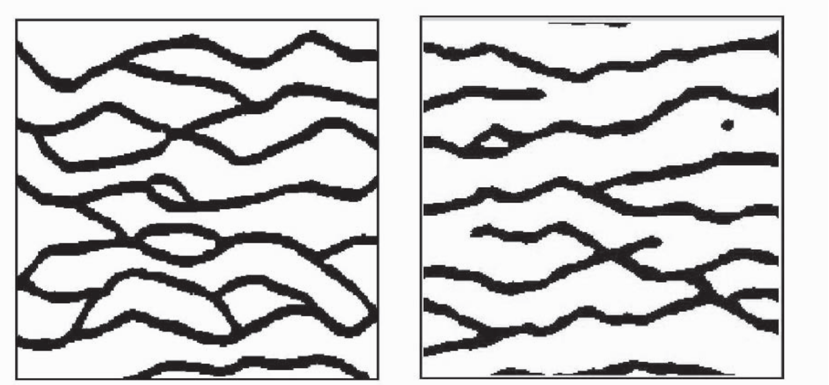


Figure 2. Training image used for the simulation of a 2D horizontal section of a fluvial reservoir (left), and reference MPS model (right).

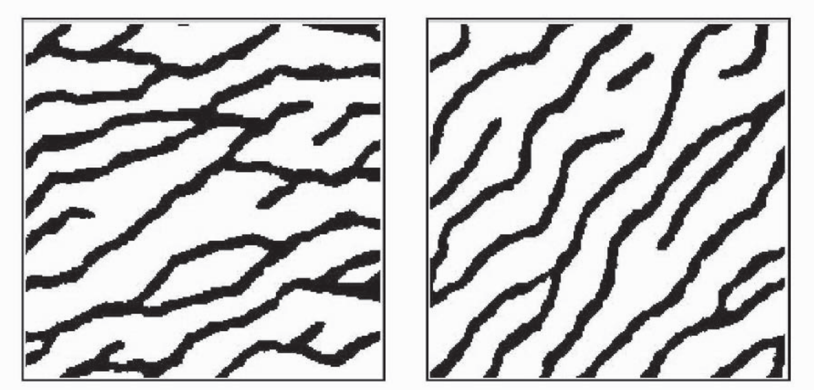


Figure 3. Fluvial reservoir MPS models obtained with a 20 degree counter-clockwise azimuth difference with the training image (left), and a 50 degree difference (right).

4.2 LOCATION-DEPENDENT AZIMUTHS

In reservoir facies modeling applications, it commonly is observed that the principal direction of continuity of the facies varies from one region of the reservoir to another. For example, topographic constraints, such as changes in the slope gradient, may lead to the formation of several sand fairways with different depositional directions. Data regarding such variations can be derived from different sources. In particular, local depositional directions can be obtained from geological interpretation (Harding *et al*, this meeting), or can be computed from seismic data (Strebelle *et al*, 2002).

Suppose that local azimuths can be estimated at each location **u** of the reservoir, and that $\theta(\mathbf{u})$ denotes the difference in degrees counter-clockwise between the azimuth value estimated at node **u** and the azimuth of the (stationary) training image. Given a data template τ_N, the method previously presented for a constant azimuth field can be extended to the location-dependent azimuth field $\theta(\mathbf{u})$ as follows:

1. Consider the range $[\theta_{min}, \theta_{max}]$ of all azimuth values estimated over the entire study field. Discretize that range into a small number L of classes, using regularly-spaced threshold values: $\theta_i = \theta_{min} + i*(\theta_{max} - \theta_{min})/L$, i=0… L.
2. Using the method described in the previous sub-section, compute for each class $[\theta_i; \theta_{i+1}]$ the search tree corresponding to the rotated data template $\tau_N(-\theta)$ where θ is the central value of the class: $\theta = (\theta_i + \theta_{i+1})/2$
3. During the simulation, at each node **u** to be simulated, use the original data template τ_N to search for nearby conditioning data, and retrieve the local cpdf from the search tree corresponding to the class of azimuth angles to which $\theta(\mathbf{u})$ belongs.

If the range $[\theta_{min}, \theta_{max}]$ of azimuth angles is greater than 90 degrees, Zhang's original technique can be used to decrease the range of the individual discretized classes $[\theta_i, \theta_{i+1}]$. For example, consider the simulation of node **u** where $\theta(\mathbf{u}) = \theta_{min} + 100°$. The rotated data template $\tau_N(90°)$ can be used to search for conditioning data (recall that 0, 90, 180, and 270 degrees are the only rotation angles for which Zhang's method requires no data relocation). Then the resulting cpdf can be retrieved from the search tree corresponding to the class of azimuth angles to which (θ_{min}+100°)-90°= θ_{min}+10° belongs. Therefore, in any case, the maximum range of azimuth values to discretize is 90 degrees. The number L of azimuth classes should depend then on the uncertainty about the local azimuth values. With L=5 classes, the range of each class is 18 degrees. This is equivalent to estimating local azimuth values with an error of ± 9 degrees.

One limitation of the above technique may be the memory demand because one search tree per azimuth class needs to be built. However, one can consider one azimuth class after the other, i.e. build the search tree corresponding to a given azimuth class, simulate all grid nodes corresponding to that class, then delete that search tree prior to considering the next azimuth class. Building, then deleting search trees is a relatively fast process compared to the actual grid simulation process.

Figure 4 shows a 2D azimuth field and a resulting simulated realization using the fluvial reservoir training image of Figure 2. The reproduction of the training patterns is similar to that in the reference simulated realization of Figure 2. Note also that, although only five azimuth classes were used, the discretization of the range of possible azimuths did not create any artifact in the simulated realization.

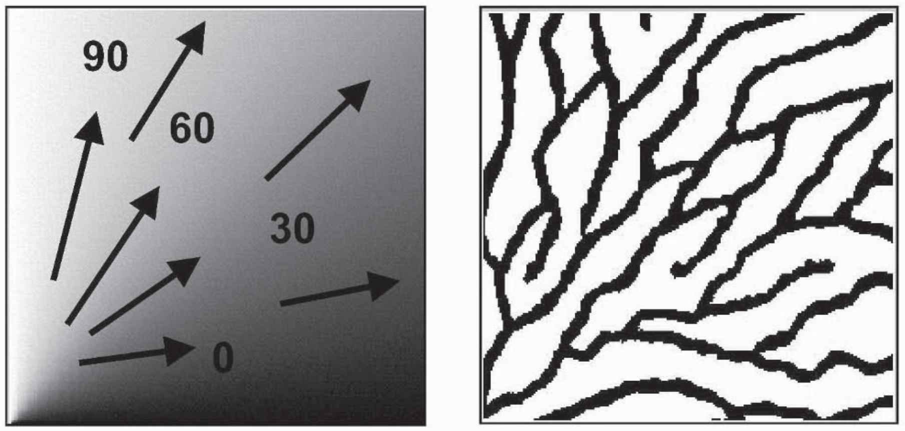


Figure 4. 2D location-dependent azimuth field (left), and resulting MPS model obtained using the fluvial reservoir training image of Figure 2 (right).

5 Integration of geobody dimensions data

Facies geobody dimensions that may depend, for example, on the distance to the sedimentation source, represent another traditional non-stationary feature of hydrocarbon reservoirs. A technique similar to that presented in the previous section to impose locally-varying azimuths is proposed to integrate geobody dimensions data, using some affinity transform of the data template used to build the search tree. For the sake of simplicity, we assume in this section that an isotropic rescaling factor (same affinity ratio in x, y, and z-directions) is sufficient to describe the variations of geobody dimensions in the volume under study.

Consider first the case in which the facies geobodies should have constant dimensions over the study field, yet possibly different from the dimensions of the training geobodies. Let λ be the ratio between target and training facies dimensions. Given a training image, and a data template $\tau_N=\{\mathbf{h}_\alpha, \alpha=1\ldots N\}$, a new data template $\tau_N(1/\lambda)$ is obtained from τ_N by the following method:

1. Rescale by $1/\lambda$ each component $\mathbf{h}_\alpha$ of τ_N. In 2D, the coordinates $(x_\alpha(1/\lambda), y_\alpha(1/\lambda))$ of the rescaled component $\mathbf{h}_\alpha(1/\lambda)$ are computed from the coordinates (x_α, y_α) of the original component $\mathbf{h}_\alpha$ as: $x_\alpha(1/\lambda)=x_\alpha/\lambda$ and: $y_\alpha(1/\lambda)=y_\alpha/\lambda$
2. Relocate these rescaled components to the nearest nodes of the training image grid.

The resulting rescaled data template $\tau_N(1/\lambda)$ is used to build the search tree from the training image. The exact same result can be obtained by rescaling the training image by λ, then building the search tree from that rescaled training image using the original data template τ_N. During the simulation, at each unsampled node, the original data template τ_N is used to search for nearby conditioning data, and the corresponding cpdf is retrieved from the above search tree.

The extension of that technique to integrate location-dependent geobody dimensions data is straightforward and similar to the integration of location-dependent azimuth data. If $\lambda(\mathbf{u})$ denotes the ratio between target and training facies dimensions at the grid node location $\mathbf{u}$, then MPS simulation using locally-varying geobody rescaling factors consists of dicretizing the range of $\lambda(\mathbf{u})$ values into a smaller number of classes, and building a search tree for the average rescaling factor value of each class.

Figure 5 shows a 2D rescaling factor field and a resulting simulated realization using the fluvial reservoir training image of Figure 2. The reproduction of the training patterns is similar to that in the reference simulated realization of Figure 2.

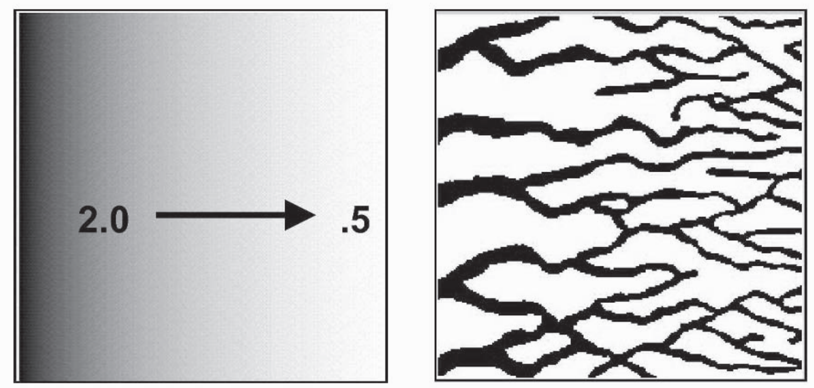


Figure 5. 2D field of location-dependent geobody dimension rescaling factors (left), and resulting MPS model obtained using the training image of Figure 2 (right).

6 Conclusion

In multiple-point geostatistics, statistics on facies joint-correlation at multiple locations are inferred from patterns displayed by a training image regardless of the location of these patterns in the training image. As a consequence, non-stationary features, such as spatial variations of facies azimuths or geobody dimensions that the training image may contain are not preserved in the multiple-point statistics simulated realizations.

To integrate variable azimuth/dimensions data, we propose applying a series of rotation/affinity transforms to a stationary training image, and building a search tree to store the multiple-point statistics inferred from each rotated/rescaled training image. During the simulation, multiple-point statistics are retrieved from the search tree corresponding to the class where the local azimuth/rescaling factor occurs. The application of that process to a 2D horizontal section of a fluvial reservoir indicates that the reproduction of the training patterns in non-stationary MPS models is similar to that observed in stationary models.

This technique can be easily generalized to create non-stationary models using several different training images thought to be representative of the geological heterogeneity in different areas of the reservoir provided that there is a smooth transition between the different training images.

References

Deutsch, C.V. and Journel, A.G., *GSLIB: Geostatistical Software Library and User's Guide*, 2nd edition, Oxford University Press, 1998.

Guardiano, F. and Srivastava, R.M., Multivariate Geostatistics: Beyond Bivariate Moments, in Soares, A., editor, *Geostatistics-Troia*, vol. 1, p. 133-144. Kluwer Academic Publications, 1993.

Holden, L., Hauge, R., Skare, Ø. and Skorstad, A., Modeling of Fluvial Reservoirs with Object Models, *Mathematical Geology*, vol. 30, no. 5, 1998, p. 473-496.

Journel, A. and Alabert, F., Non-Gaussian Data Expansion in the Earth Sciences, *Terra Nova* , no. 1, 1989, p. 123-134.

Lia, O., Tjelmeland, H. and Kjellesvik, L.E., Modeling of Facies Architecture by Marked Point Models, in *Geostatistics-Wollongong*, p. 386-398, Kluwer Academic Publications, 1997.

Strebelle, S., *Sequential Simulation Drawing Structures from Training Images*, Ph.D. Thesis, Department of Geological and Environmental Sciences, Stanford University, 2000.

Strebelle, S., and Journel, A., Reservoir Modeling Using Multiple-point Statistics, paper SPE 71324 presented at the 2001 SPE Annual Technical Conference and Exhibition, New Orleans, Sept. 30-Oct. 3, 2001.

Strebelle, S., K. Payrazyan, and J. Caers, Modeling of a Deepwater Turbidite Reservoir Conditional to Seismic Data Using Multiple-Point Geostatistics: paper SPE 77425 presented at the 2002 SPE Annual Technical Conference and Exhibition, San Antonio, Sept.29-Oct. 2, 2002.

Viseur, S., Stochastic Boolean Simulation of Fluvial Deposits: a New Approach Combining Accuracy and Efficiency, paper SPE 56688 presented at the 1999 SPE Annual Technical Conference and Exhibition, Houston, Oct. 3-6, 1999.

Zhang, T. Program 2001 snesim version 5.0. In *Report 15, Stanford Center for Reservoir Forecasting*, Stanford, CA, 2002.

A WORKFLOW FOR MULTIPLE-POINT GEOSTATISTICAL SIMULATION

YUHONG LIU
ExxonMobil Upstream Research Company, P. 0. BOX 2189, Houston, TX 77252, USA. E-mail: yuhong@pangea.stanford.edu

ANDREW HARDING, RUSTY GILBERT
ChevronTexaco Energy Technology Company, San Ramon, CA 94583, USA

ANDRE JOURNEL
Geological and Environmental Science Department, Stanford University, Stanford, CA 94305, USA

Abstract. There are presently two main avenues in the stochastic modeling of depositional facies: pixel-based and object-based geostatistics. They both have strengths and weaknesses: traditional pixel-based geostatistics is good at data conditioning, but it depends on variograms to capture spatial structures and hence fails to reproduce definite patterns common to most geological facies; while object-based geostatistics is good at reproducing crisp facies shapes but is difficult to condition to dense well data or exhaustive 3D seismic data. Multiple-point simulation, a newly developed pixel-based technique, integrates the strengths of both: it keeps the flexibility of pixel-based techniques for data conditioning, while allowing pattern reproduction through consideration of multiple-point statistics. In this paper, a workflow for multiple-point stochastic simulation is discussed in details. This workflow is applied to an industry project. The results show reproduction of the prior geological knowledge and honoring of both well and seismic data.

1 Introduction

Integrating all available information when building a geological model is a recurrent and difficult problem. One challenge lies in how to condition the model to different types of measured reservoir data, such as wells and 3D seismic data. Another challenge lies in how to account for prior geological knowledge, a fuzzy yet important conceptual information. Geostatistics provides an ensemble of tools for data integration and uncertainty evaluation (Deutsch and Journel, 1992; Goovaerts, 1997). Through computer-based data integration algorithms, multiple equi-probable numerical models of the reservoir properties are built, whose difference reflects uncertainty.

There are two main avenues in geostatistical modeling: pixel-based and object-based. Object-based geostatistics performs simulation by "dropping" different geological bodies one after another onto the simulation field (Deutsch and Wang, 1996). Any

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 245-254.

added geological body is accepted, rejected, or modified through evaluating some objective function measuring the match to local data. Hence by its nature, it is good at reproducing the crisp shapes of geological bodies. But often it is CPU demanding, particularly when extensive hard data must be honored; also their capability in integrating 3D seismic data is limited: typically only 2D aerial proportion maps derived from seismic can be accounted for. By contrast, pixel-based algorithms simulate each grid node of the reservoir model one pixel at a time (Deutsch and Journel, 1992; Goovaerts, 1997). All the unsampled nodes are sequentially visited along a random path. The probability distribution function (cpdf) at any given node is estimated conditional to all data (both hard and soft) found in its neighborhood. A value is drawn from that cpdf using Monte Carlo simulation. This simulated value is frozen as hard data for simulation of the subsequent unknown nodes. Because pixel-based geostatistics performs simulation one pixel at a time, it is very flexible and easy to condition to most conditional data. However, the traditional pixel-based geostatistics uses the variogram, a two-point statistic, to capture the spatial structures of facies. It has been found that it is difficult for a simple variogram model to capture the curvilinear structures of geological bodies. Higher order statistics are required. For example, Figure 1 shows three distinct images with similar variograms (Strebelle, 2000).

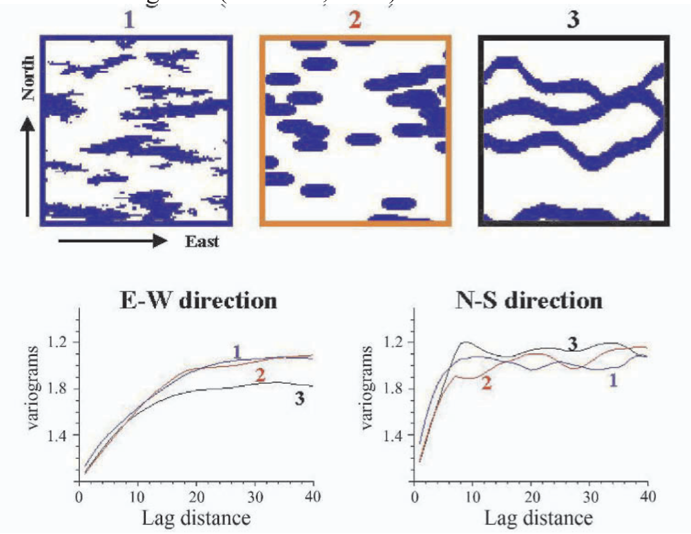


Figure 1. Three distinct images with similar variograms. (Source: Strebelle, 2000, slightly modified.)

Hence none of the traditional geostatistical simulation techniques are ideal for depositional facies modeling integrating both geology and 3D seismic data. Multiple-point geostatistics (Journel, 1992; Guardiano and Srivastava, 1993; Strebelle, 2000; Strebelle and Journel, 2000; Strebelle, et al, 2002; Liu, 2003; Liu, et al, 2004) is called upon to address this problem. In the following sections two case studies are presented to illustrate the application of multiple-point geostatistics. Next, a workflow of multiple-point simulation is proposed.

2 Why multiple-point geostatistics?

Multiple-point geostatistics aims at reproducing complex statistics involving two and more points at a time. This allows capturing information much beyond the reach of a mere variogram model. Being an advanced pixel-based technique, multiple-point geostatistics inherits the advantage of building the numerical model one pixel at a time, allowing easy data conditioning. Yet compared with the traditional variogram-based algorithms, it has the enhanced capability of reproducing curvilinear shapes of geological bodies, a feature traditionally reserved to object-based algorithms. Two examples are presented in this section to respectively illustrate these two strengths.

2.1 BETTER INTEGRATION OF GEOLOGY

This first example is a synthetic case study of building a numerical reservoir property model from limited well data. Figure 2a shows a photo of the Wagon Caves Rock outcrop (Anderson et al, 1998). From this outcrop, rock properties such as sand/shale indicators, grain size, porosity and permeability, are measured along the two marked vertical columns. They are taken as known well data, while rock properties at all other grid nodes are assumed unknown and are simulated using different geostatistical algorithms.

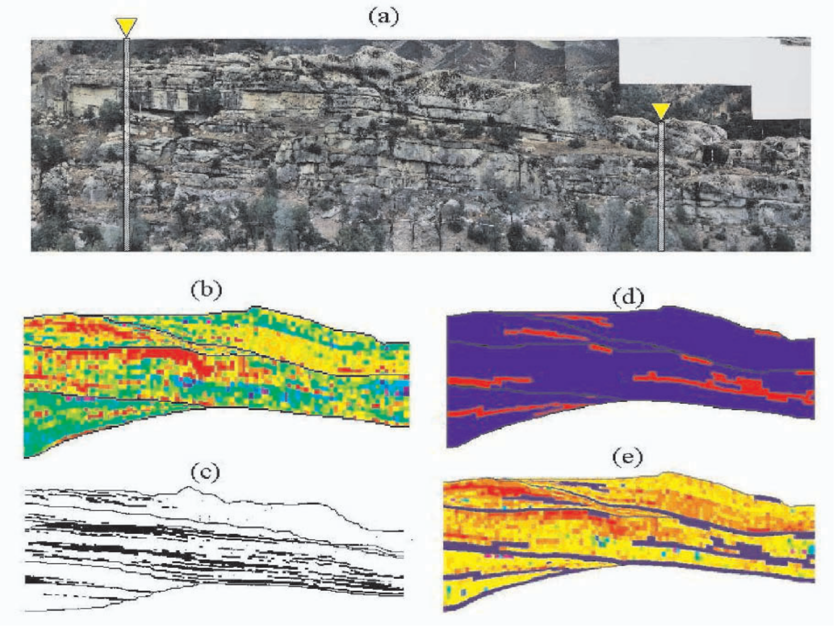


Figure 2. Integration of geology using a multiple-point simulation algorithm. (a) the Wagon Rock Caves outcrop, from which, two vertical columns are taken as well data; (b) one realization of permeability by the two-point model; (c) training image of mud layers, used for multiple-point simulation; (d) one multiple-point realization of mud layers; (e) one permeability realization including simulated mud layers.

First a variogram-based two-point geostatistical simulation is performed. Figure 2b shows one realization of permeability using sequential Gaussian simulation (Deutsch and Journel, 1998): it displays high-entropy spatial patterns typical of Gaussian simulations. One major problem with any two-point model is that it can not reproduce the elongated thin mud layers, which can act as flow barriers due to their very low permeabilities.

From the outcrop, two types of mud layers are observed: continuous mud layers spanning across the two wells and discontinuous mud layers that pinch out before reaching either one of the two wells. Multiple-point simulation is performed to reproduce these different types of mud layers. Figure 2c depicts the shapes of the mud layers over a vertical section, which can serve as a training image for multiple-point simulation of these mud layers. For sensitivity analysis purpose, we build three models: one without mud layers, one with discontinuous (pinched-out) mud layers, and one with continuous mud layers crossing all the way between the two wells. Figure 2d shows one realization of the simulated mud layers using the multiple-point simulation *snesim* algorithm (Strebelle, 2000). Figure 2e shows one realization of permeability including the simulated mud layers.

To analyze the flow response of these different models, we perform a single-phase steady-state upscaling over the whole model to get a single upscaled permeability tensor for the whole simulation field. A high effective permeability along a certain direction means easy flow along that direction. Figure 3a shows the effective permeability along the vertical direction (denoted as Kz) for three different variogram-based models, each model with a different range and represented by 10 equi-probable realizations. When the range is decreased from 1200 to only 150, Kz changes within a small range of 140-220 md. In contrast, when a multiple-point algorithm is used to incorporate different types of mud layers, Kz changes dramatically within a range of 220-30-1 md (Figure 3b).

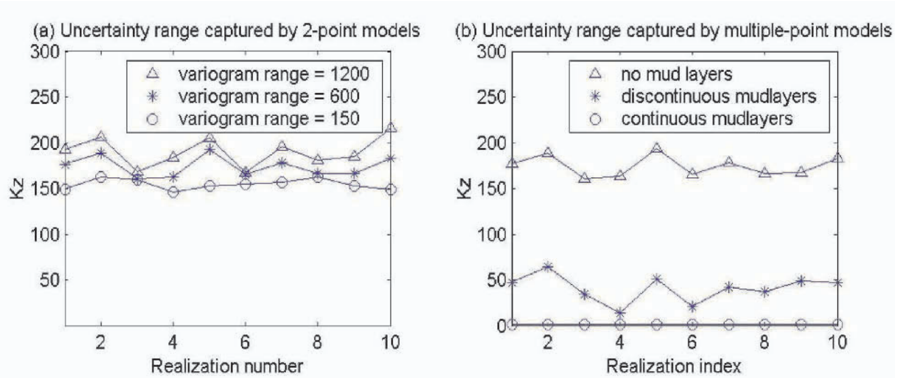


Figure 3. Vertical effective permeability (Kz) of different models. (a) Kz of three different variogram-based models; (b) Kz of three different multiple-point models.

This example illustrates the importance of simulating correctly the crisp shapes and continuity of geological bodies: they can have significant impact on flow response, a major concern of petroleum engineers. Traditional variogram-based geostatistics fails in this aspect, while multiple-point geostatistics achieves it at little additional CPU cost, provided a conceptual training image of the type of Figure 2c is available.

2.2 EASY DATA CONDITIONING

Another advantage of multiple-point simulation is easier data conditioning. Object-based algorithms can reproduce crisp shapes and continuity of geological bodies. A major problem with object-based algorithms, however, is difficult data conditioning, especially in presence of dense well data and diverse types of soft data, such as 3D seismic data or production data. Figure 4 illustrates such a real reservoir study (Liu, 2003; Liu, et al, 2004).

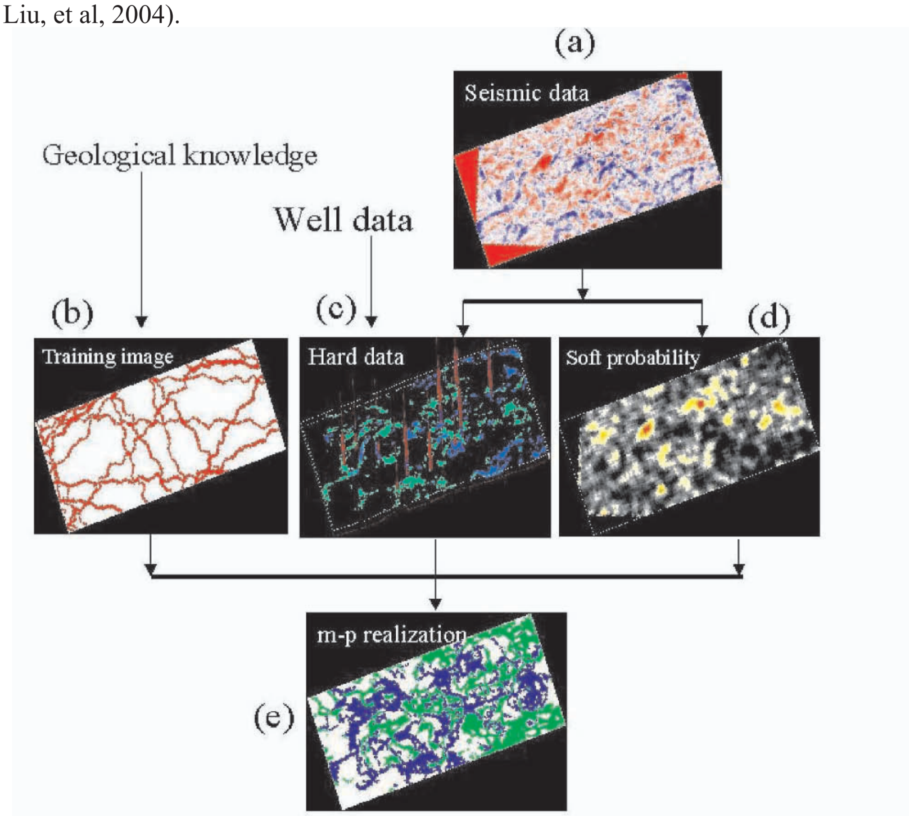


Figure 4. Multiple-point simulation integrating diverse types of information (Note all the following data/information/realizations are in 3D, although only horizontal slices are shown here). (a) seismic data; (b) a training image depicting the prior geological concepts; (c) hard data (well data + seismic imaged channel pieces); (d) seismic-derived soft probability for sand; (e) one multiple-point realization honoring the information shown in (b), (c), (d).

In this reservoir characterization study, there are four different sources of information:

- Geological knowledge: the depositional environment of this reservoir is interpreted as a fluvial channel system, which can be summarized by a training image (see Figure 4b).
- Well data: they are considered as hard data and must be reproduced exactly by all simulated realizations (see the vertical columns in Figure 4c).
- Geobodies extracted from seismic data: good quality 3D seismic data can clearly image some characteristic geobodies (channel segments in this case study). These geobodies (see the two types of clusters in Figure 4c) are deemed certain and need to be reproduced exactly by all realizations. A PCA clustering technique (Scheevel and Payrazyan, 1999), capable of recognizing such characteristic facies pieces, is used to extract these geobodies from the 3D seismic data.
- Soft information from seismic data: in areas not clearly imaged by seismic data, a soft probability data cube for presence of sand can be derived from the seismic data (see Figure 4d). All alternative simulated models for sand/shale should be all consistent with the seismic data in that probabilistic sense.

While it would be extremely difficult to constrain object-based simulation to all these different sources of information, multiple-point simulation can easily achieve this, because it operates one pixel at a time. Figure 4e presents one such a multiple-point realization.

3 A workflow for multiple-point simulation

In multiple-point geostatistics, a training image is used to deliver prior geological concepts about the geometry of reservoir heterogeneities. This training image should be reasonably stationary, and deliver the shapes, patterns and distributions of geological objects deemed present in the actual reservoir. It essentially plays the same role as a variogram model in traditional two-point geostatistics: it provides statistics relating the unsampled value to conditioning data involving jointly multiple locations. The simulation process amounts to take the training image patterns, "morphing" and anchoring them to location-specific reservoir data.

Strebelle (2000) developed a *snesim* algorithm, which significantly speeds up the original multiple-point simulation algorithm proposed by Guardiano and Srivastava (1993). In this program, the training image is scanned only once to retrieve the frequency of occurrences of observed outcomes for the central nodal value given a template of neighboring conditioning data. These probabilities are then stored into a search tree data structure, which allows fast storage and retrieval of probabilities corresponding to the actual hard conditioning data events encountered during sequential simulation. Conditioning to hard sample data (e.g. well data) in multiple-point simulation is done the same way as in two-point algorithms. The hard data values are frozen at their nodal locations and never changed; each unknown node is sequentially visited and simulated conditional to the original hard data and previously simulated values. As for soft data conditioning (e.g. seismic data or production data), a Bayesian-type paradigm is applied in the multiple-point simulation workflow. Instead of using some cokriging algorithms as in two-point geostatistics (Goovaerts, 1997), soft data conditioning is performed in two steps. The first step is to extract the useful information from the soft data. For example, a prior facies conditional probability is derived from the seismic data. The second step is to perform simulation integrating hard and soft data, which can be further decomposed into three sub-steps. At each unknown node, first the probability conditioned to the current multiple-point hard data event is read from the search tree. Then it is combined with the previously derived soft conditional probability to get the final or posterior probability conditioned to both hard and soft data. Finally, a facies indicator value is drawn from that posterior probability.

This multiple-point simulation workflow can be subdivided into three parts, each of which has a different conditional probability involved (Figure 5). All three parts are later explained in more detail.

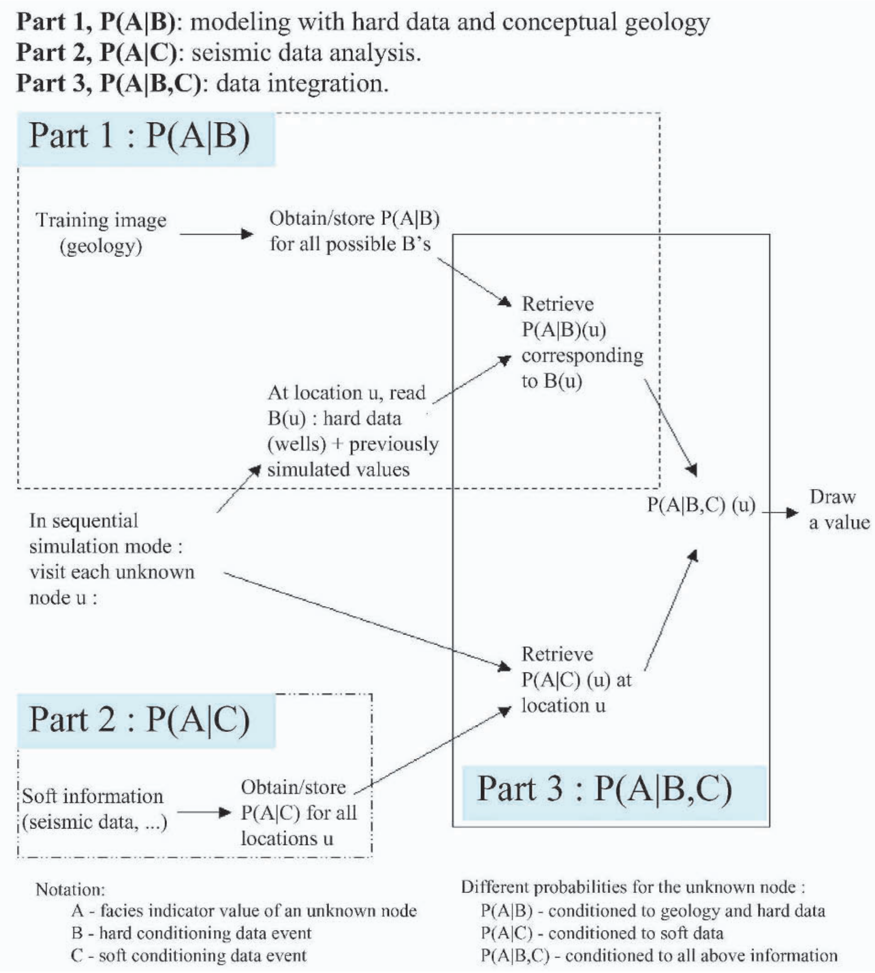

Part 1, P(A|B): modeling with hard data and conceptual geology
Part 2, P(A|C): seismic data analysis.
Part 3, P(A|B,C): data integration.
Part 1 : P(A|B)
Training image (geology)
Obtain/store P(A|B) for all possible B's
Retrieve P(A|B)(u) corresponding to B(u)
At location u, read B(u) : hard data (wells) + previously simulated values
In sequential simulation mode : visit each unknown node u :
P(A|B,C) (u)
Draw a value
Retrieve P(A|C) (u) at location u
Part 2 : P(A|C)
Soft information (seismic data, ...)
Obtain/store P(A|C) for all locations u
Part 3 : P(A|B,C)
Notation:
A - facies indicator value of an unknown node
B - hard conditioning data event
C - soft conditioning data event
Different probabilities for the unknown node :
P(A|B) - conditioned to geology and hard data
P(A|C) - conditioned to soft data
P(A|B,C) - conditioned to all above information

Figure 5. A multiple-point simulation workflow, decomposed into three parts.

3.1 P(A|B): MODELING WITH HARD DATA AND CONCEPTUAL GEOLOGY

P(A|B) denotes the conditional probability of the value to be simulated given a multiple-point hard conditioning data event B, with A representing, e.g., a facies indicator value. This part aims at capturing and reproducing the geological information provided by the training image, conditional to the hard data event B.

First a data template, composed of multiple nodes with any user-specified configuration, is used to scan the training image, and the number of replicates of each different multiple-point data event is retrieved. These numbers are stored in a search tree data structure (Strebelle, 2000), which allows an easy retrieval of information.

Next, in a pixel-based simulation mode, each uninformed node **u** is sequentially visited. Its neighboring conditioning data event B (including both the original hard sample data and previously simulated values) is collected. Two numbers are then retrieved from the search tree: number of replicates of the joint data event (A,B) and the number of replicates of the conditioning data event B. The multiple-point conditional probability P(A|B) is then easily calculated as:

$$P(A\mid B)=\frac{P(A,B)}{P(B)}=\frac{\text{number of replicates for }\mathit{(A,B)}}{\text{number of replicates for }\mathit{(B)}} \tag{1}$$

This conditional probability P(A|B) is used to either directly draw a value for the node **u**, if no soft information is available, or is combined with any co-located soft data conditional probability P(A|C) to get a joint conditional probability P(A|B,C). The value at location **u** is then drawn using the updated probability P(A|B,C), see hereafter.

3.2 P(A|C): SEISMIC DATA ANALYSIS

As discussed above, when there is soft information, such as seismic data, it is necessary to determine the conditional probability P(A|C), denoting the facies probability given the soft data C (say, seismic) alone. This part tries to establish the relationship between seismic patterns, and geological patterns, enabling prediction of rock properties from the measured seismic data. Many different techniques can be used to retrieve this probabilistic information from the seismic data. They can be divided into two categories: supervised vs. unsupervised techniques. Supervised techniques are used when there exist calibration geological data associated with the seismic data, for example, well data versus corresponding seismic data, or interpreted geological facies versus corresponding seismic data. The pattern recognition from seismic data is then "supervised" by the known geological data. In contrast, unsupervised techniques try to directly identify patterns from seismic data without any prior geological constraints. Techniques in both categories have been used by geostatisticians: supervised or unsupervised neural network (Caers, 1999), principal components clustering (Scheevel and Payrazyan, 1999; Strebelle et al., 2002; Liu, 2003), maximum message length technique (Arroyo, 2000), the latter uses entropy to measure the dispersion of different seismic patterns.

3.3 P(A|B,C): DATA INTEGRATION

After retrieving useful information separately from the geology + hard data, and from seismic, that is, obtaining the two individual conditional probabilities P(A|B) and P(A|C), the next step is to combine them into one single posterior probability, P(A|B,C), conditioned to all available information. This is the data integration part.

Journel (2002) proposed a "Permanence of Updating Ratios" paradigm to integrate P(A|B) and P(A|C) into P(A|B,C). The basic assumption of this algorithm is that the relative contribution of data event C is the same before and after knowing B:

$$\frac{x}{b} = \frac{c}{a} \tag{2}$$

where, a, b, c and x represent distances to the event A occurring defined as:

$$a = \frac{1 - P(A)}{P(A)}, \quad b = \frac{1 - P(A \mid B)}{P(A \mid B)}, \quad c = \frac{1 - P(A \mid C)}{P(A \mid C)}, \quad x = \frac{1 - P(A \mid B,C)}{P(A \mid B,C)}$$

All these distances are bounded within [0,∞). They reach 0 if the probability of A occurring is 1, infinity if that probability is 0.

From the permanence relation (Eq.2), P(A|B,C) is calculated as:

$$P(A \mid B,C) = \frac{1}{1 + x}$$

Zhang and Journel (2003) later showed that the previous permanence assumption is equivalent to a Bayesian updating under conditional independence of B and C given A. To account for dependence between B and C data, Journel proposed the following generalization using a power parameter $\tau > 0$:

$$\frac{x}{b} = \left(\frac{c}{a}\right)^{\tau} \tag{3}$$

Setting $\tau > 1$ increases the impact of seismic data, conversely, setting $\tau < 1$ decreases the impact of seismic data.

These three parts establish a general workflow for multiple-point geostatistical simulation.

4 Conclusions

In this paper, a multiple-point simulation workflow is proposed and discussed. Anchored in the pixel-based category, which allows an easier conditioning to a variety of data, the multiple-point approach aims at identifying and reproducing the spatial patterns typically displayed by geological bodies. Hence it incorporates the advantages of both pixel-based techniques and object-based techniques. The proposed multiple-point simulation workflow is composed of three parts:

- Modeling with hard data and conceptual geology, i.e., obtaining P(A|B): The prior geological knowledge is represented by a training image, which is scanned to obtain the probability P(A|B), namely, the probability of presence/absence of a facies A given its multiple-point hard conditioning data event B.

- Seismic data analysis, i.e., obtaining P(A|C): This part establishes the relationship between seismic patterns (C) and facies patterns (A). The result is a probability P(A|C) field, denoting the facies probability given the neighboring multiple-point seismic data C.
- Integration of different sources of information, i.e., obtaining P(A|B,C): This part integrates the two previous individual conditional probabilities, P(A|B) and P(A|C), into one single posterior probability P(A|B,C), conditioned to both geology and seismic data. The facies indicator (A) at each unsampled node is drawn from this updated posterior probability.

Acknowledgements

We would like to thank ChevronTexaco Energy Technology Company and its affiliates for the permission to publish one of the case studies. The helps provided by Sebastien Strebelle, William Abriel and staffs of ChevronTexaco are especially acknowledged.

References

Anderson, K.; Hickson, Thomas A.; Crider, G and Graham, S.: Integrating teaching with field research in the Wagon Rock Project. Journal of Geoscience Education, vol.47, no.3, May 1999, pp.227-235.

Caers, J.: Modeling Facies Distributions from Seismic Using Neural Nets. SCRF Annual Report No.13, vol.1. 2000.

Deutsch, C.V. and Journel, A.G., GSLIB: Geostatistical Software Library and User's Guide, Oxford University Press, 1992.

Deutsch, C. and Wang, L.: Hierarchical Object-Based Stochastic Modeling of Fluvial Reservoirs. Mathematical Geology, vol. 28, no. 7, 1996, p 857-880.

Gilbert, R., Liu, Y., Abriel, W. and Preece, R.: Reservoir Modeling Integrating Various Data and at Appropriate Scales, The Leading Edge, Vol.23, no. 8, Aug. 2004, p784-788.

Goovaerts, P., Geostatistics for Natural Resources Evaluation, Oxford University Press, 1997.

Guardiano, F. and Srivastava, R.M.: "Multivariate Geostatistics: Beyond Bivariate Moments", Geostatistics-Troia, A. Soares (ed.), Kluwer Academic Publications, Dordrecht, 1993, vol 1, p 113-114.

Journel, A: Geostatistics: Roadblocks and Challenges. In A. Soares (Ed.), Geostatistics-Troia, Kluwer Academic Publ., Dordrech, 1992, p 213-224.

Journel, A: Combining knowledge from diverse sources: an alternative to traditional data independence hypotheses, Mathematical Geology, vol. 34, no. 5, 2002.

Liu, Y.: Downscaling Seismic Data into A Geological Sound Numerical Model, Ph.D. dissertation, Department of Geological and Environmental Science, Stanford University, Stanford, 2003, pp. 202.

Liu, Y., Harding, A., Abriel, W. and Strebelle, S.: Multiple-Point Simulation Integrating Wells, 3D Seismic Data and Geology, AAPG Bulletin, Vol. 88, No.7, 2004.

Scheevel, J. R., and Payrazyan, K.: Principal Component Analysis Applied to 3D Seismic Data for Reservoir Property Estimation, SPE 56734, SPE Annual Technical Conference and Exhibition, Houston, 1999

Srivastava, M., 1995: An Overview of Stochastic Methods for Reservoir Characterization, in Yarus, J., and Chambers, R., eds., Stochastic modeling and geostatistics: principles, methods, and case studies, v.3: AAPG Computer Applications in Geology, p. 3-16.

Strebelle, S.: Sequential Simulation Drawing Structures from Training Images, Ph.D. dissertation, Department of Geological and Environmental Sciences, Stanford University, Stanford, 2000.

Strebelle, S., and Journel, A.: Reservoir Modeling Using Multiple-point Statistics, SPE 71324, SPE Annual Technical Conference and Exhibition, 2001.

Strebelle, S., Payrazyan, K. and Caers, J.: Modeling of a Deepwater Turbidite Reservoir Conditional to Seismic Data Using Multiple-Point Geostatistics, SPE 77425, SPE Annual Technical Conference and Exhibition, San Antonio, Texas, 2002.

Zhang, T. and Journel, A. G.: Merging Prior Geological Structure and Local Data: the mp Geostatistics Answer, SCRF Annual Report No.16, vol. 2, 2003.

A MULTIPLE-SCALE, PATTERN-BASED APPROACH TO SEQUENTIAL SIMULATION

G. BURC ARPAT and JEF CAERS
Department of Petroleum Engineering, Stanford University
367 Panama St., Stanford, CA 94305-2220, USA

Abstract. In the context of multiple-point geostatistics, a new algorithm (SIMPAT) is presented. This algorithm relies on several image processing concepts, such as image similarity, to borrow and reproduce patterns from training images constrained to hard and soft data. The method makes use of a new multiple-grid approach by which the scale relations between the training image patterns are better captured and reproduced.

1 Introduction

Sequential simulation is one of the most widely used stochastic imaging techniques within the Earth Sciences. The theory is well understood (Daly, 2004; Goovaerts, 1997) and many practical, fast and robust algorithms have been developed (Deutsch and Journel, 1998) such as sequential Gaussian simulation (SGSIM) and sequential indicator simulation (SISIM).

However, realizations generated by SGSIM (and also SISIM) are often deemed too 'synthetic' looking, not reflecting the actual variability of Earth Science phenomena such as facies distributions in oil reservoirs or sedimentary deposits in aquifer systems. The limitations of SGSIM lie in the assumption of a multi-Gaussian distribution that requires knowledge of a histogram and a variogram. The variogram, as a two-point statistics, is not capable of modeling complex, connected and curvilinear spatial variation. To overcome these limitations, multiple-point geostatistics (MPS) was introduced together with the concept of "training image" (Guardiano and Srivastava, 1993). A training image is an exhaustive 3D picture containing patterns believed to be similar to the actual field under investigation. The training image serves as a concept, a vision of what spatial variability of the study area should look like. As a mere concept, the training image need not be constrained to any hard or soft data.

Based on the original MPS idea, Strebelle (2000) proposed a practical algorithm (SNESIM) that generates realizations mimicking the 3D patterns of the training image while constraining to hard and soft data. The algorithm works in the same way as many other sequential algorithms do: (1) visit each node of the simulation grid randomly; (2) at each node, estimate the conditional probability given the neighboring data and previously simulated nodes (called a "data event"), and (3) draw from that probability

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 255-264.

distribution and assign the value to the node. In SNESIM, the conditional probability is sampled from the training image by looking for replicates of the data event.

In this paper, an alternative pattern-based algorithm is proposed by redefining the problem of pattern reproduction as an image processing problem. In image processing, one generally tackles complex images by finding common patterns in the image and working on these patterns (Palmer, 1999). A similar approach can be devised for geostatistical modeling where one finds all the patterns of a training image. These patterns correspond to multiple-pixel configurations within a user-defined template and capture meaningful pieces of geological shapes known to exist in field of study. Such patterns exist at different geological scales and patterns at various scales interact with each other. The idea is to generate realizations that reproduce these multiple-scale patterns on the simulation grid. A new practical algorithm SIMPAT (SIMulation with PATterns) is implemented to achieve this goal. The paper shortly describes the inner workings of this algorithm, presents some 3D examples and discusses how SIMPAT complements the already existing sampling-based algorithms such as SNESIM.

2 A New, Pattern-based Sequential Simulation Method

2.1 NOTATION

Define $z(\mathbf{u})$ as the realization of a random variable $Z(\mathbf{u})$ modeling the variable of study where $\mathbf{u} = (x,y,z) \in \mathbf{G}$ and $\mathbf{G}$ is the regular Cartesian grid discretizing the field of study. $Z(\mathbf{u})$ can be a model of either a continuous or a categorical variable. The random function itself is denoted as $\mathbf{Z} = \{Z(\mathbf{u}), \forall \mathbf{u} \in \text{study area}\}$ and a realization as $\mathbf{z}$.

$\mathbf{z_T}(\mathbf{u})$ indicates a location-specific vector of $z(\mathbf{u})$ within a template $\mathbf{T}$ centered at $\mathbf{u}$, i.e.:

$$\mathbf{z_T}(\mathbf{u}) = \{z(\mathbf{u}+\mathbf{h}_0), z(\mathbf{u}+\mathbf{h}_1), \ldots, z(\mathbf{u}+\mathbf{h}_\alpha), \ldots, z(\mathbf{u}+\mathbf{h}_{n_\mathbf{T}-1})\} \qquad (1)$$

where $\mathbf{h}_\alpha$ vectors are the vectors defining the geometry of the $n_\mathbf{T}$ nodes of the template $\mathbf{T}$ and $\alpha = 0, \ldots, n_\mathbf{T} - 1$ with the special vector $\mathbf{h}_0 = 0$ identifying the node $\mathbf{u}$. A flag notation $z(\mathbf{u}) = \chi$ is used for 'unknown' nodes, i.e. nodes still to be informed by the sequential simulation and hence that do not have an assigned value yet.

To distinguish the training image, the hard and the soft data from the simulated realization $\mathbf{z}$, the notations **ti**, **hd** and **sd** are used. For example, a multiple-point event scanned from the training image **ti** at location $\mathbf{u}'$ is denoted by $\mathbf{ti_T}(\mathbf{u}')$, i.e.:

$$\mathbf{ti_T}(\mathbf{u}') = \{ti(\mathbf{u}'+\mathbf{h}_0), ti(\mathbf{u}'+\mathbf{h}_1), \ldots, ti(\mathbf{u}'+\mathbf{h}_\alpha), \ldots, ti(\mathbf{u}'+\mathbf{h}_{n_\mathbf{T}-1})\} \qquad (2)$$

where location $\mathbf{u}' \in \mathbf{G}'$ and $\mathbf{G}'$ is the regular Cartesian grid discretizing the training image. The training image grid $\mathbf{G}'$ need not be the same as the realization grid $\mathbf{G}$.

A pattern $\mathbf{pat_T}^k$ is the particular k-th configuration of the above vector of values $\mathbf{ti_T}(\mathbf{u}')$ defined by the template $\mathbf{T}$ where $k = 0, \ldots, n_\mathbf{pat} - 1$ and $n_\mathbf{pat}$ is the number of total available patterns. Each k-th configuration is assumed to be location-independent and, thus, the vector $\mathbf{pat_T}^k$ is written as:

$$\mathbf{pat}_{\mathbf{T}}^{k} = \left\{ pat_{\mathbf{T}}^{k}(\mathbf{h}_0), pat_{\mathbf{T}}^{k}(\mathbf{h}_1), \ldots, pat_{\mathbf{T}}^{k}(\mathbf{h}_\alpha), \ldots, pat_{\mathbf{T}}^{k}(\mathbf{h}_{n_{\mathbf{T}}-1}) \right\} \qquad (3)$$

where all patterns are defined on the same template **T**.

In sequential simulation, a data event $\mathbf{dev}_{\mathbf{T}}(\mathbf{u})$ is defined as the set of hard data and previously simulated values neighboring the visited location **u** within the template **T**, i.e. $\mathbf{dev}_{\mathbf{T}}(\mathbf{u}) = \mathbf{z}_{\mathbf{T}}(\mathbf{u})$.

The dissimilarity (distance) between a data event and a pattern is calculated using a node-based distance function:

$$d\left\langle \mathbf{dev}_{\mathbf{T}}(\mathbf{u}), \mathbf{pat}_{\mathbf{T}}^{k} \right\rangle = \sum_{\alpha=0}^{n_{\mathbf{T}}-1} \left| dev_{\mathbf{T}}(\mathbf{u} + \mathbf{h}_\alpha) - pat_{\mathbf{T}}^{k}(\mathbf{h}_\alpha) \right| \qquad (4)$$

where *d*<> denotes the distance function (Manhattan distance; Duda et al., 2001). When for a certain node $\mathbf{u} + \mathbf{h}_\alpha$, $dev_{\mathbf{T}}(\mathbf{u} + \mathbf{h}_\alpha,) = \chi$ (unknown), the value is ignored in the distance calculation. For other distance functions that can be used with the SIMPAT algorithm, the reader is referred to Arpat (2004).

2.2 THE SINGLE-GRID, UNCONDITIONAL SIMPAT ALGORITHM

The algorithm starts by scanning the training image using a template **T** to acquire all patterns of **ti**. A filter can be applied to discard undesirable patterns. Remaining patterns are stored in a pattern database and such patterns are denoted by $\mathbf{pat}_{\mathbf{T}}^{k}$ where the size of the pattern database is $n_{\mathbf{pat}}$ as defined in the previous section.

The simulation part of the algorithm follows the sequential simulation framework. During simulation, nodes are randomly visited and the data event $\mathbf{dev}_{\mathbf{T}}(\mathbf{u})$ is extracted. Then, $\mathbf{dev}_{\mathbf{T}}(\mathbf{u})$ is compared to all available patterns in the pattern database using a predefined similarity criterion. The aim is to find the 'most similar' pattern to the data event, denoted by $\mathbf{pat}_{\mathbf{T}}^{*}$. In other words, the algorithm minimizes *d*<> of Equation 4 for all patterns $\mathbf{pat}_{\mathbf{T}}^{k}$ and labels the minimum as $\mathbf{pat}_{\mathbf{T}}^{*}$. Once this most similar pattern is found, the data event $\mathbf{dev}_{\mathbf{T}}(\mathbf{u})$ is replaced by $\mathbf{pat}_{\mathbf{T}}^{*}$, i.e. the values of $\mathbf{pat}_{\mathbf{T}}^{*}$ are pasted on to the simulation grid at the current node **u**.

The above outlined algorithm can be divided into two main parts:

(1) Pre-processing of the training image:

P-1. Scan the training image using the template **T** to obtain all existing patterns $\mathbf{pat}_{\mathbf{T}}^{k}$ that occur over the training image.

P-2. Reduce the number of patterns to $n_{\mathbf{pat}}$ by applying filters to construct the pattern database. Typically, only unique patterns are taken, i.e. repetitions (frequency) of patterns are ignored.

(2) Simulation on the simulation grid:

S-1. Define a random path on the simulation grid to visit each node **u** only once.

S-2. At each node **u**, retain the data event $\mathbf{dev}_{\mathbf{T}}(\mathbf{u})$ and find the $\mathbf{pat}_{\mathbf{T}}^{*}$ that minimizes $d\!<\mathbf{dev}_{\mathbf{T}}(\mathbf{u}), \mathbf{pat}_{\mathbf{T}}^{k}>$ for $k = 0, \ldots, n_{\mathbf{pat}} -1$, i.e. $\mathbf{pat}_{\mathbf{T}}^{*}$ is the 'most similar' pattern.

S-3. Once the most similar pattern $\mathbf{pat_T}^*$ is found, assign $\mathbf{pat_T}^*$ to $\mathbf{dev_T}(\mathbf{u})$, i.e. for all the $n_\mathbf{T}$ nodes $\mathbf{u} + \mathbf{h}_\alpha$ within the template $\mathbf{T}$, $dev_\mathbf{T}(\mathbf{u} + \mathbf{h}_\alpha) = pat_\mathbf{T}^*(\mathbf{h}_\alpha)$.

S-4. Move to the next node of the random path and repeat the above steps until all the grid nodes along the random path are exhausted.

On large simulation grids, a practical problem occurs due to the finite size of the template **T**. To capture the large scale correlations of the training image, a large template would need to be used. Figure 5b of Section 4 (Examples) demonstrates this problem. Yet, using a large template would make the minimization step (Step S-2) of the SIMPAT algorithm too CPU demanding. To overcome this problem, SIMPAT employs a modified version of the multiple-grid approach as proposed by Tran (1994). The idea is to use a set of cascading multiple-grids and sparse templates instead of a single grid and one large dense template. The simulation is first performed on the coarse grid and then these coarse values are passed to the subsequent finer grids as conditioning information. This idea is elaborated below.

2.3 THE MULTIPLE-GRID, UNCONDITIONAL SIMPAT ALGORITHM

On a Cartesian grid, the multiple-grid view of a grid **G** is defined by a set of cascading coarse grids $\mathbf{G}^g$ and templates $\mathbf{T}^g$ instead of a single fine grid and one large dense template where $g = 0, \ldots, n_g - 1$ and n_g is the total number of multiple-grids. The g-th coarse grid ($0 \leq g \leq n_g - 1$) is constituted by each 2^g-th node of the final grid ($g = 0$) in each direction. If **T** is a template defined by vectors $\mathbf{h}_\alpha$, then the template used for a coarser grid $\mathbf{T}^g$ is defined by $\mathbf{h}_\alpha^g = 2^g \times \mathbf{h}_\alpha$ and has the same configuration of $n_\mathbf{T}$ nodes as **T** but with spacing 2^g times larger. Figure 1 illustrates this concept.

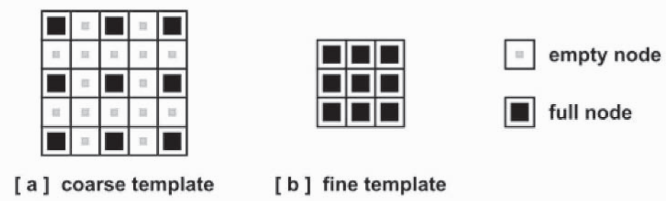


Figure 1. A 3x3 fine template (b) and its corresponding coarse template (a) obtained by expanding the fine template with 2^g spacing where $g = 1$ (the first coarse grid).

The multiple-grid simulation of a realization is achieved by successively applying the single-grid algorithm explained above to the multiple-grids starting from the coarsest grid. After each multiple-grid simulation, the values calculated on the current grid are transferred to the one finer grid and g is set to $g - 1$. This succession of multiple-grid simulations continues until $g = 0$. On a multiple-grid, the previously calculated coarser grid values contribute to the distance calculation of Step S-2, i.e. if on node **u** a previous coarse grid value exists, this value is taken into account when minimizing the distance.

Different from the classical multiple-grid approach (Tran, 1994), SIMPAT does not 'freeze' the coarse grid values when they are transferred to a finer grid, i.e. such values are still allowed to be updated and visited by the algorithm in the subsequent multiple-grid simulations. In other words, the coarse grid nodes are always included in the finer grid simulations.

The above multiple-grid approach allows the values determined on the coarser grids to be modified by the finer grids, i.e. coarse to fine scale interaction. For complex training images, fine to coarse interaction might also be desired to fully capture the scale relations of training image patterns. SIMPAT utilizes a feedback mechanism, termed "dual template simulation", that allows such fine to coarse scale interaction.

Consider the pattern $\mathbf{pat}_{\mathbf{T}}^{k}$ scanned from the training image at location **u** using the template **T** for a coarse grid simulation. Another, fine template **T′** can be used at the same location to obtain the corresponding fine pattern such that template **T′** covers the same area (volume in 3D) as template **T** but with all the nodes of the finest grid. **T′** is called the "dual template" of **T**. The relation between **T** and **T′** is shown in Figure 2.

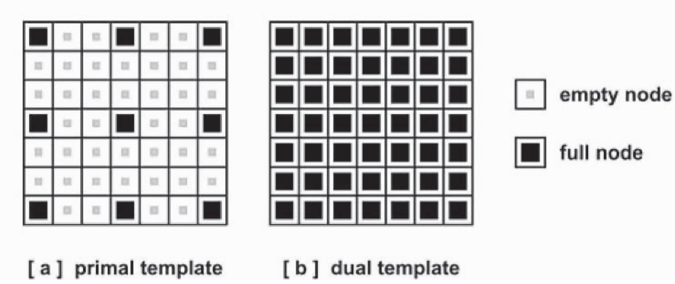


Figure 2. Illustration of the primal (a) and the dual (b) template concepts.

The two patterns scanned using **T** and **T′** (called the "primal pattern" and the "dual pattern") are linked to each other in the pattern database. Then, during the simulation, whenever a coarse $\mathbf{pat}_{\mathbf{T}}\mathbf{*}$ is found using the distance calculations and pasted on to the coarse simulation grid, the corresponding fine pattern (scanned from the same location **u** but with template **T′**) retrieved from the pattern database is simultaneously pasted on to the fine grid. In essence, the multiple-grid simulation of the realization is performed in parallel on all grids but using only the similarity criterion of the current coarse grid. Figure 3 illustrates the steps of this approach for a single node **u** on the coarse grid of a 2 multiple-grid simulation.

The values simulated using the dual templates will affect the results of the subsequent distance calculations on the finer grids, thus allowing the desired feedback from the finer grids to the coarse grids. Consider the case of 3 multiple grids. When the coarsest grid (g = 2) simulation is completed, due to the use of the dual templates, the simulation grid will be completely full. Then, during the middle grid simulation, the values previously pasted by the dual template on to the finest grid will affect the distance calculations of the middle grid, hence providing the fine to coarse feedback. In essence, the algorithm places the large scale patterns on the coarsest grid and 'roughly' decides on small scale patterns and then 'corrects' for finer details on the subsequent grids.

Figure 5d of Section 4 (Examples) demonstrates the effect of the modified multiple-grid approach as used in SIMPAT.

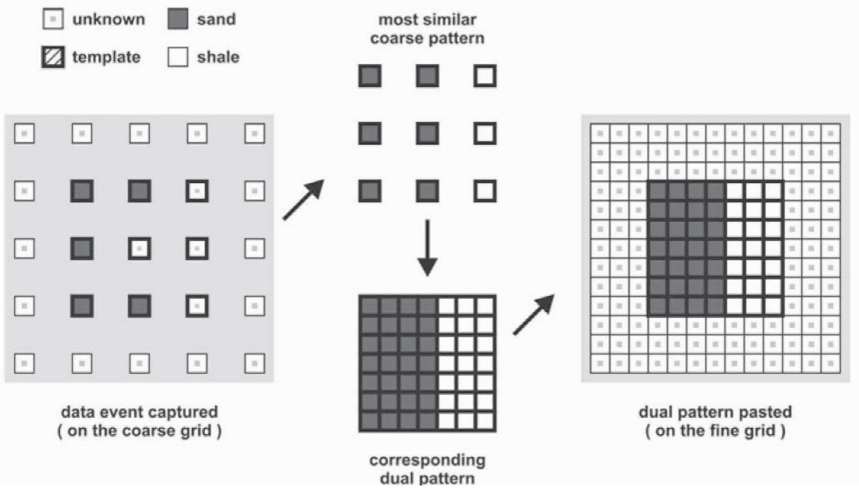

unknown
sand
template
shale
most similar coarse pattern
data event captured (on the coarse grid)
corresponding dual pattern
dual pattern pasted (on the fine grid)

Figure 3. Illustration of using dual templates with a binary (sand/shale) variable. First, the data event $\mathbf{dev}_\mathbf{T}(\mathbf{u})$ on the coarse grid is captured. The most similar pattern $\mathbf{pat}_\mathbf{T}*$ to this data event is found by minimizing the distance between $\mathbf{dev}_\mathbf{T}(\mathbf{u})$ and $\mathbf{pat}_\mathbf{T}^k$. Then, the corresponding dual pattern of $\mathbf{pat}_\mathbf{T}*$ is retrieved from the pattern database and pasted on to the finest grid.

3 Data Conditioning

3.1 HARD DATA CONDITIONING

In SIMPAT, conditioning to hard data is performed in Step S-2 of the algorithm, during the search for the most similar pattern. If conditioning data exists on any $dev_\mathbf{T}(\mathbf{u} + \mathbf{h}_\alpha)$, the algorithm first checks whether $pat_\mathbf{T}^k(\mathbf{h}_\alpha)$ is equal to this data. If the pattern $\mathbf{pat}_\mathbf{T}^k$ does not fulfill this condition (i.e. there is a mismatch), it is skipped and the algorithm searches for the next most similar pattern until a match is found. If none of the available patterns fulfill the condition, the algorithm selects a pattern such that only the nodes of the data event that has conditioning information are considered during the distance calculations and other nodes are ignored. If several patterns fulfill this condition, then a second minimization is performed on the non-conditioning nodes of the data event using only these patterns. In essence, a two-stage similarity check is performed: first, only for the conditioning data; then, for the previously simulated nodes.

3.2 SOFT DATA CONDITIONING

In Earth Sciences, soft data typically refers to data obtained from indirect measurements acquired using some form of remote sensing (e.g. geophysical methods). Thus, soft data is nothing but a ‘filtered’ view of the original field of study, where the filtering is performed by some forward model **F**. In general, the forward model **F** is not known exactly and is approximated by a known model **F***.

In SIMPAT, conditioning to soft data calls for a soft training image. This soft training image can be obtained by applying the approximate forward model **F*** to the (hard) training image (See Figure 7c and 7d of Section 4 for an example). The patterns of the soft data (for example, a response from a seismic survey) are related to the geological (hard) patterns of the realization through the above mentioned filter model. The pair of hard and soft training images provides a basis for modeling the multiple-point relationship between hard and soft patterns. In fact, any joint statistics or pattern pairs extracted from the two training images can be considered as the multiple-point alternative of a cross-variogram in variogram-based geostatistics.

Once the soft training image is obtained, SIMPAT explicitly relates the patterns in the hard training image and the soft training image by creating a joint pattern database. In other words, Step P-1 of the pre-processing part of the algorithm is modified such that, for every $\mathbf{pat}_\mathbf{T}^k$ of the hard training image, a corresponding soft pattern is extracted from the soft training image from the same location **u**. Another modification is done to the Step S-2 of the simulation part of the algorithm, i.e. the search for the most similar pattern. Instead of minimizing the distance between $\mathbf{dev}_\mathbf{T}(\mathbf{u})$ and $\mathbf{pat}_\mathbf{T}^k$, the algorithm now minimizes,

$$d^{1,2}\langle\cdot,\cdot\rangle = \omega\times d^1\left\langle \mathbf{dev}_\mathbf{T}(\mathbf{u}),\mathbf{pat}_\mathbf{T}^k\right\rangle + (1-\omega)\times d^2\left\langle \mathbf{dev}_\mathbf{T}^2(\mathbf{u}),\mathbf{pat}_\mathbf{T}^{2,k}\right\rangle \qquad (5)$$

i.e., the summation of two distances where $\mathbf{dev}_\mathbf{T}^2(\mathbf{u})$ denotes the soft data event obtained from the soft data grid **sd**, $\mathbf{pat}_\mathbf{T}^{2,k}$ is a soft pattern and ω is a weight that is attached to the combined summation to let the user of the algorithm give more weight to either the hard or the soft values, reflecting the 'trust' of the user to the soft data. The flowchart of these modifications when conditioning only to soft data is given in Figure 4.

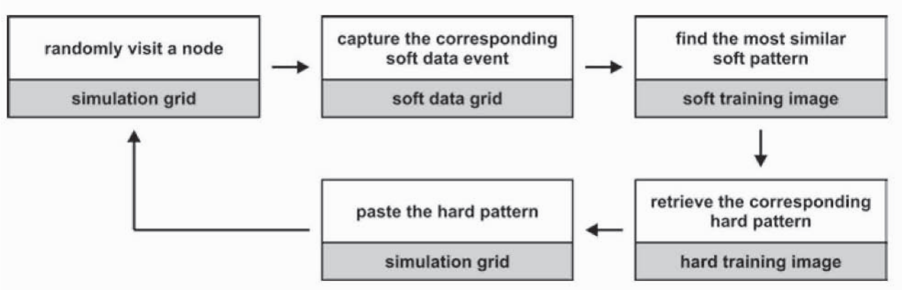


Figure 4. The flowchart for the conditional search for the most similar pattern $\mathbf{pat}_\mathbf{T}$* when the hard data event is not informed. When there is conditioning data or previously simulated nodes within the hard data event, a joint search is performed instead.

The net result of the above modifications is that, for every node **u**, the algorithm now finds the most similar pattern not only based on the previously calculated nodes but also based on the soft data. When there is also hard data available, this minimization is performed only after the patterns that condition to the hard data are found as explained in the previous section, i.e. hard data has priority over soft data.

4 Examples

Figure 5a is a 7 facies training image depicting a tidal channel system in an oil reservoir. A notable property of Figure 5a is that, the image is highly non-stationary, especially the large scale variation: note how one facies appears only on the front part of the cube. Figure 5b is an unconditional SIMPAT realization obtained using a single-grid simulation. Figure 5c shows the application of the traditional multiple-grid approach (Tran, 1994). Section 2.3 explains two modifications done to this traditional approach where (1) the coarse grid values are not 'frozen' on the finer grids and (2) the dual template simulation technique is employed. Figure 5d is an unconditional SIMPAT realization obtained using these modifications. As these final figure illustrate, the algorithm successfully captures the non-stationary behavior of the training image, while adequately reproducing the facies relations of the training image.

Figure 6a shows a synthetic reference case with 6 facies in an oil reservoir. A dense data set is sampled from this reference to test conditioning to hard data (Figure 6b). The training image used is shown in Figure 6c. The final conditional SIMPAT realization is in Figure 6d. The training image used in this example is highly representative of the reference case; both the reference and the training image contain stacked channels. This agreement keeps the number of conflicting patterns to a minimum during the simulation. For a more realistic case, the reader is referred to Arpat (2004).

Figure 7 demonstrates the application of soft data conditioning using SIMPAT. In this case, soft data is obtained by applying a seismic forward model **F*** to the binary reference case (Wu, Mukerji and Journel, 2004). The same model is applied to the training image to obtain the soft training image. The final SIMPAT realization (Figure 7f) conditions to soft data relatively well but pattern reproduction is somewhat degraded as made evident by the disconnected channel pieces. This issue, along with possible solutions, is further discussed in Arpat (2004).

5 Conclusion

The sequential simulation method of SIMPAT replaces the traditional probability framework of drawing from conditional probability distributions (for example, as used in the SNESIM algorithm of Strebelle, 2000) with calculations of similarity between patterns. This entirely new approach to stochastic simulation has the advantage that it focuses directly on one of the core purposes of stochastic simulation: reproduction of patterns (Be it two-point or multiple-point, stationary or non-stationary). The similarity approach does not share many of the restrictions of a probabilistic approach, which often calls for a rather strong assumption of stationarity in the inference or modeling of patterns via probabilities. While the initial results are promising, the downside of this approach is that the purely algorithmic formulation of the method to stochastic pattern reproduction and data conditioning is not yet well understood. Future research will therefore focus on understanding better the advantages and limitations of the various new concepts (similarity, dual templates, similarity-based data conditioning, etc.) presented in this paper.

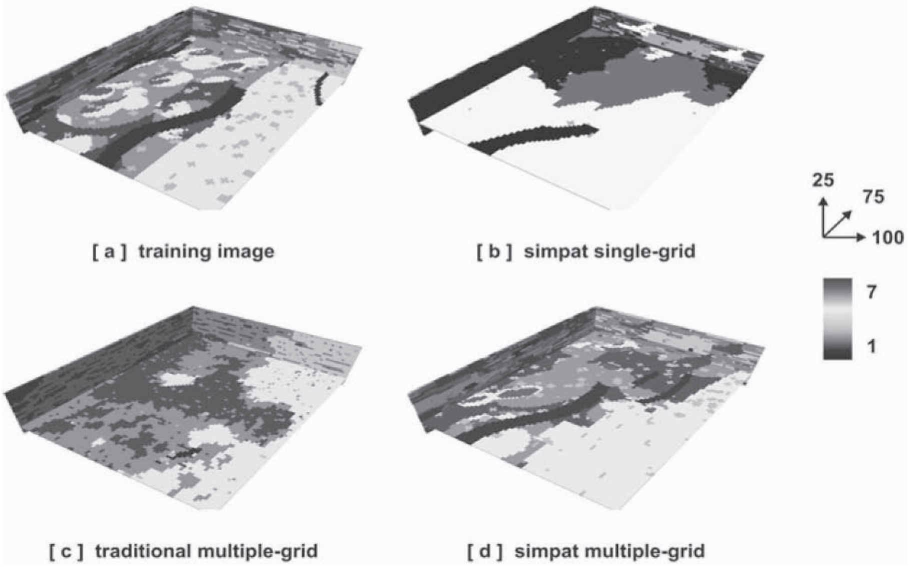


Figure 5. Unconditional SIMPAT. (b) - (d) all use 11×11×5 templates. In (b), only a single grid is used. (c) utilizes the traditional multiple-grid method with 3 multiple-grids (where simulated nodes are frozen and stay constant for the rest of the simulation) and (d) is obtained using the new multiple-grid approach that employs the dual template simulation technique.

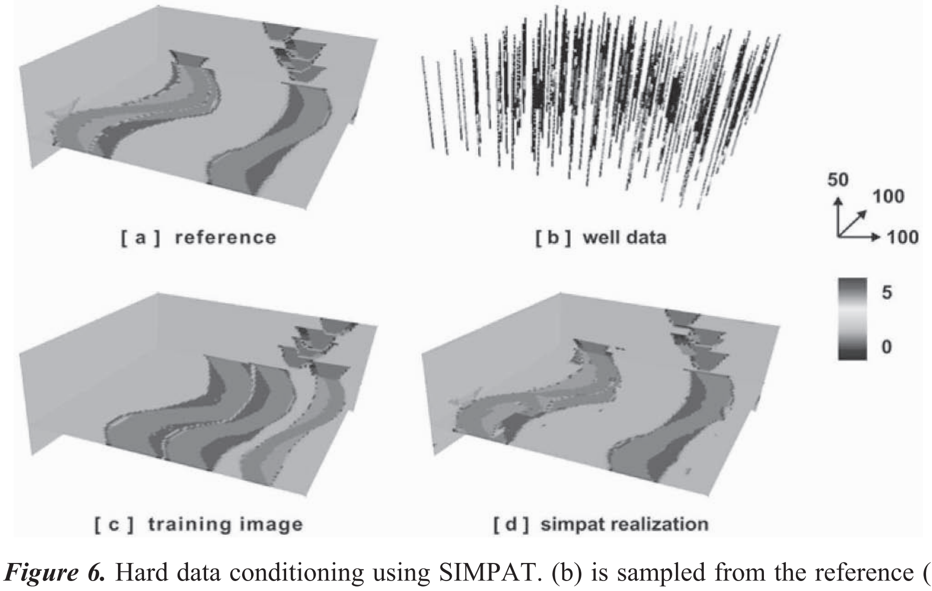


Figure 6. Hard data conditioning using SIMPAT. (b) is sampled from the reference (a) and constitutes 2% of all nodes. (c) is the training image and (d) is the final conditional SIMPAT realization obtained using a 11×11×5 template and 3 multiple-grids.

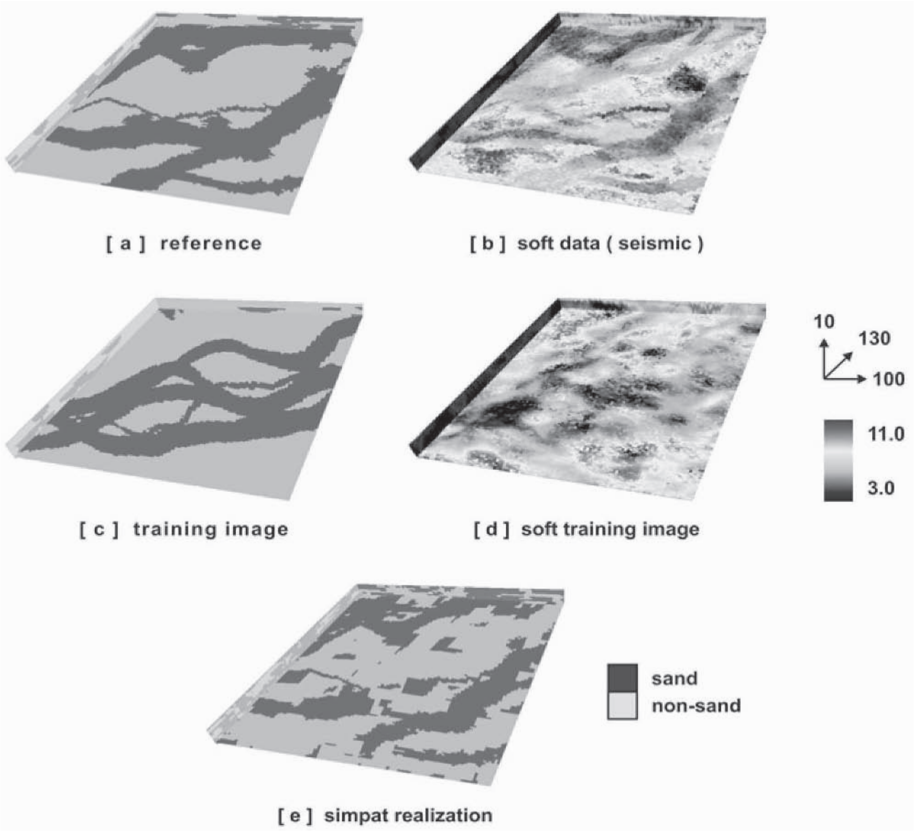


Figure 7. Soft data conditioning using SIMPAT. The soft training image (d) is obtained by applying an approximate model **F*** to (c). The final conditional SIMPAT realization (e) is obtained using a 11×11×3 template and 3 multiple-grids.

References

Arpat, G. B., SIMPAT: stochastic simulation with patterns, 17th SCRF Meeting, Stanford Center for Reservoir Forecasting, Stanford University, 2004.

Daly, C., Higher order models using entropy, Markov random fields and sequential simulation, *Geostatistics-Banff*, Kluwer Academic Publications, 2005.

Deutsch, C. and Journel, A., *GSLIB: Geostatistical Software Library*, Oxford University Press, 2nd ed., 1998.

Duda, O., Hart, P. and Stork, D., *Pattern Classification*, John Wiley & Sons, Inc., 2nd ed., 2001.

Goovaerts, P., *Geostatistics for Natural Resources Evaluation*, Oxford University Press, 1997.

Guardiano, F. and Srivastava, R., Multivariate geostatistics: Beyond bivariate moments, *Geostatistics-Troia*, Kluwer Academic Publications, 1993, pp. 133 - 144.

Palmer, S.E., *Vision Science: Photons to Phenomenology*, MIT Press, 1999.

Strebelle, S., *Sequential Simulation Drawing Structures from Training Images*, Ph.D. thesis, Stanford University, 2000.

Tran, T., Improving variogram reproduction on dense simulation grids, *Computers and Geosciences*, vol. 20 no. 7, 1994, pp. 1161–1168.

Wu, J., Mukerji, T. and Journel, A., Prediction of spatial patterns of saturation time-lapse from time-lapse seismic, *Geostatistics-Banff*, Kluwer Academic Publications, 2005.

SEQUENTIAL CONDITIONAL SIMULATION USING CLASSIFICATION OF LOCAL TRAINING PATTERNS

T. Zhang, P. Switzer, A. Journel
Department of Geological and Environmental Sciences
Stanford University, CA, 94305, U.S.A

Abstract Local spatial structures, as depicted by a training image, can be summarized by a few general linear filter scores. Local training patterns are then classified according to these scores. Sequential simulation proceeds by associating each conditioning multiple-point data event with a score class and then patching a pattern from this class onto the simulation grid. This procedure can handle both binary and continuous variable training images as illustrated by several diverse training images.

1 Introduction

Multiple point (mp) simulation aims to capture local patterns from a training image (TI) and anchor them to actual data. A training image reflects only general aspects of spatial structure or texture. It should display stationary patterns, which can be transported to the actual simulation space, see Figure 1a for an horizontal 2D fault training image example.

The original mp simulation concept was introduced by Srivastava, in Guardiano and Srivastava (1993). The original algorithm was very CPU-demanding in that the training image had to be rescanned for each node being simulated. Strebelle (2002) traded the CPU problem for a greater RAM demand by scanning the TI only once and storing all required information in a search tree data structure. Strebelle's program *snesim* made mp simulation feasible for 3D applications but limited to the joint simulation of no more than 4 or 5 categories.

In this paper, we propose a new mp simulation approach that can deal with both categorical and continuous variable training images with reasonable CPU and memory demand. Dimension reduction, hence RAM saving, is obtained by classifying the training patterns according to a few linear filters. A neighborhood template is passed over the training image. At each pixel location of the training image the template records the local pattern as an array of values. The array is reduced to a low-dimensional set of scores by applying a few general linear filters, see Figure 2. Patterns with similar scores are then grouped together into pattern classes.

To capture patterns at different scales, the template can be rescaled to scan the same training image. For example, if the finest template comprises N pixels at horizontal spacing 1x1, a coarser template with the same topology would comprise again N pixels

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 265-273.

but with spacing 2x2, or 4x4. This corresponds to the concept of multiple grids commonly used in sequential simulation (Tran, 1994).

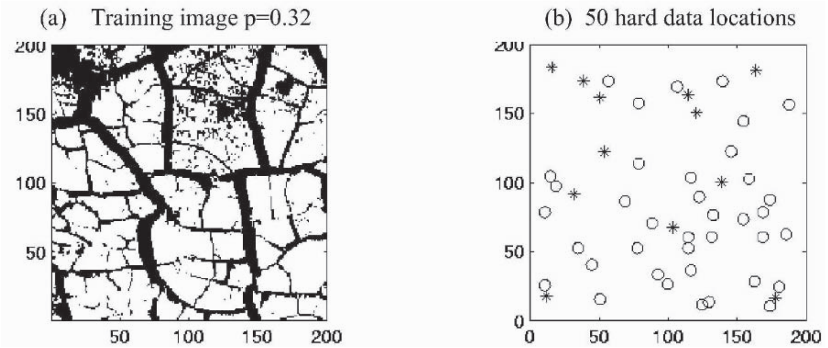


Figure 1. (a) training image with p=32% of faults, (b) 50 data locations sampled from (a) (star-faults; circle-background)

2 Pattern scoring

Let $X(i,j)$ denote the value at location (i,j) in the training image. A score $S_f(i,j)$ for the pattern in the neighborhood of (i,j) is defined for a filter $f(u,v)$ as follows:

$$S_f(i,j) = \sum_{v=-n}^{n} \sum_{u=-n}^{n} X(i+u, j+v) f(u,v)$$

where the dimension of the local neighborhood or template is (2n+1)x(2n+1).
We define six different filters $f_1, ..., f_6$ as follows.

(1) f_1 **:** N-S average

$$f_1(u,v) = 1 - \frac{|v|}{n}, v = -n, ..., n$$

see Figure 2a.

(2) f_2 **:** E-W average

$$f_2(u,v) = 1 - \frac{|u|}{n}$$

see Figure 2b.

(3) f_3 : N-S gradient

$$f_3(u,v) = \frac{v}{n}$$

see Figure 2c.

(4) f_4 : E-W gradient

$$f_4(u,v) = \frac{u}{n}$$

see Figure 2d.

(5) f_5 : N-S curvature

$$f_5(u,v) = \frac{2|v|}{n} - 1$$

see Figure 2e.

(6) f_6 : E-W curvature

$$f_6(u,v) = \frac{2|u|}{n} - 1$$

see Figure 2f.

Each of these six filters is used to scan the TI. At each pixel location, the template of neighborhood data is weighted by the filters to produce a series of 6 scores. If the six scores are assigned to the pixel at the center of the template, we thus obtain score maps of the training image itself. In Figure 3, we see the score maps for the training image in Figure 1a. The size of the template used is 27x27 pixels (n=13), while the training image is 200x200.

The first two score maps S_1 and S_2 are weighted moving averages of the 27x27=729 template values. They highlight the object center locations. The next two scores S_3 and S_4 come from gradient filters; they provide edge detection, and highlight the object boundary contrast. The last two scores S_5 and S_6 are derived by curvature filters, they provide gradient changes. Note that these 6 filters privilege the NS and EW directions; appropriate rotations should be applied to either the training image or the filter weight maps if one wishes to emphasize different directions yet with the same total number of filters.

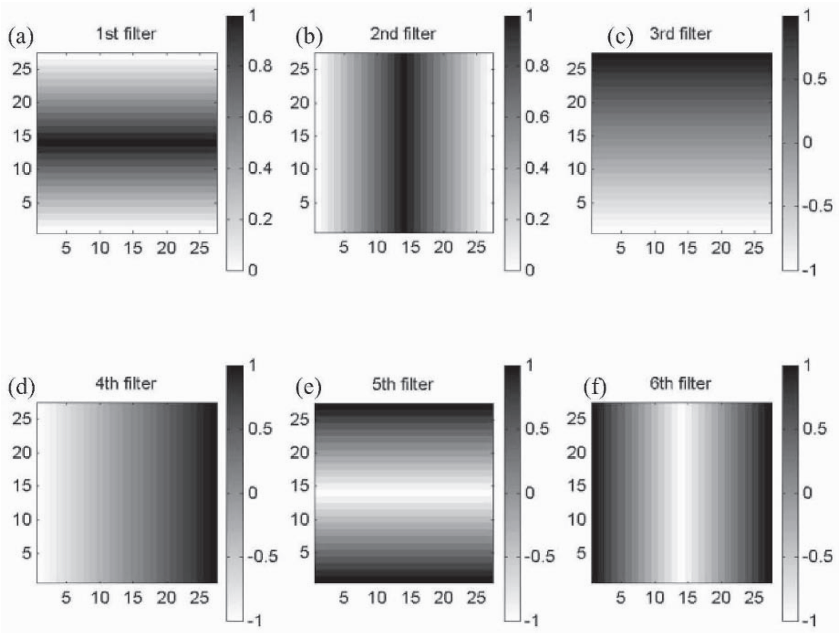


Figure 2. Six general filters (27x27)

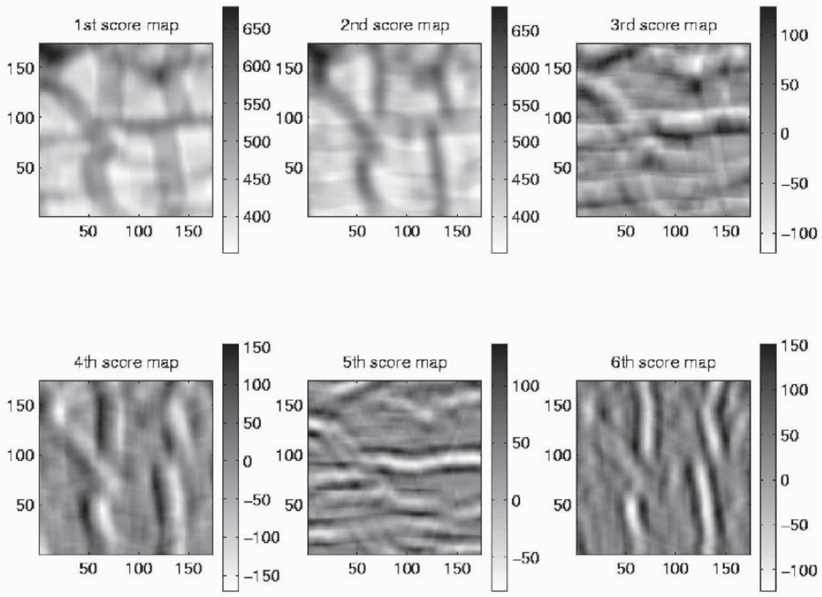


Figure 3. Six score maps at the finest grid

3 Pattern classification

Scanning the TI with each of the six filters produces a frequency distribution for each of the scores $f_1, ..., f_6$. Each of these marginal frequency distributions is discretized into 5 equal frequency bins according to their respective quintiles. This results in a partition of the 6-dimensional score space into $5^6 = 15625$ cells. [For binary data on a fine grid it may happen that many templates consist of all zeros or all ones. Therefore it is possible for some quintiles to be the same, resulting in fewer effective bins].

Even though each of the six scores has been divided into equal frequency bins, the 6-component joint cell frequencies are not equal. Many cells are empty because there are no local training patterns having such filter score combinations. Training patterns whose filter scores fall into the same cell are thus grouped into pattern classes. For each non-empty score cell, a “prototype” is obtained by averaging all patterns falling into that class, which can be seen as the aggregate of similar training patterns. Figure 4 shows the first 8 prototypes with the most training pattern replicates taken from Figure 1a on the finest scale with a template size 27x27.

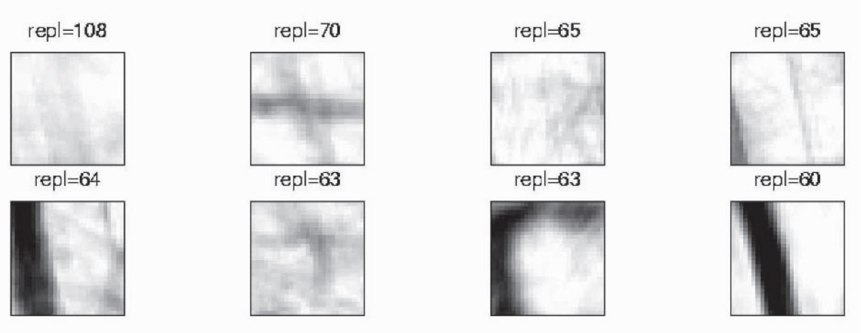


Figure 4. 8 prototypes with the most replicates from the training image of Fig. 1a at the finest grid (27x27)

4 Pattern simulation

Based on the previous classification of local training patterns, sequential simulation with multiple-grids can be utilized to generate pattern simulations that together mimic structural features of the training image.

At each node to be simulated, conditioning data are searched within a data template centered at this node. This data template has the same dimensions as that used to scan the training image at the current grid level.

If there are no conditioning data within the data template, we choose the template prototype closest to the target global mean value and pick a training pattern from this prototype class, the pattern whose mean is closest to the target mean. A target mean value is specified before the simulation and it is expected that the averaged value of each realization should be close to this value. If there are conditioning data in the data template, calculate the distance between this data event (DEV) and each training prototype (PROT) template recorded at the current grid level.

There are three types of conditioning data, $k = 1,2,3$:
(1) hard original data
(2) previously frozen simulated nodes
(3) non-frozen previously simulated nodes from pattern patches, see below.
The distance expression is written as

$$d(DEV, PROT) = \sum_{k=1}^{3} \omega(k) \frac{\sum_{i_k=1}^{n_k} |x^k(i_k) - y^{(k)}(i_k)|}{n_k}$$

where i_k are the pixel locations of information of type k and $\omega(k)$ are weights for the three respective information types with $\omega(1) \geq \omega(2) \geq \omega(3)$.

Once we identify the local template prototype closest to the conditioning template information, we sample a specific pattern from the prototype pattern class. We patch the sampled pattern at the current simulation node but retaining hard data and previously frozen simulated locations. The "inner" part of the patch is frozen and is not revisited in the sequential simulation. A larger inner patch area makes simulation faster, but may cause discontinuity. The outer part of the patch will be revisited, hence re-simulated. The concept of using a patch instead of a single node can be found in texture synthesis (Liang et al., 2001).

We use multi-grids to capture pattern structure and texture at different scales. The training image is scanned using local templates at several grid scales. Separately, at each grid scale, local patterns are converted to 6-dimensional scores using the filters described earlier. Thus, at each grid scale we get a classification of local patterns at that scale. In our examples, we used two coarser scales that are 2 times and 4 times the dimension of the finest grid scale.

Simulation proceeds from the coarsest grid to the finest grid. Simulated values from the preceding coarser grid are used as conditioning information at the finer grid simulation. However, all coarser grid simulated values are revisited and re-simulated at the finer grid.

5 Illustrations

The illustrations shown here exhibit the ability of local filter scores to capture spatial patterns when used to simulate categorical and continuous patterns from training images. First, we investigate fault structural simulation using the 200x200 binary categorical training image of Figure 1a. The area proportion covered by faults in the training image is p=0.32.

Three grid scales are used; the local template size is 27x27 pixels; the patch size is 19x19 pixels. 50 hard data, as shown in Figure 1b, are sampled from the training image. After training image classification, sequential simulation proceeds by randomly visiting grid nodes (pixels) and identifying local pattern prototypes that match the currently available information in the neighborhood of the simulation node. The target proportion of fault area was set to 0.30. Figures 5a-5c display the same conditional simulated realization at the different grid scales, the final simulated image being at the right of the figure. Figure 6 displays three additional conditional simulations at the final grid. It can be seen that the fault structures, including small fractures, are reasonably reproduced, although with less large scale continuity than displayed on the training image of Figure 1a. The 50 hard conditioning data are honored exactly by all simulated realizations.

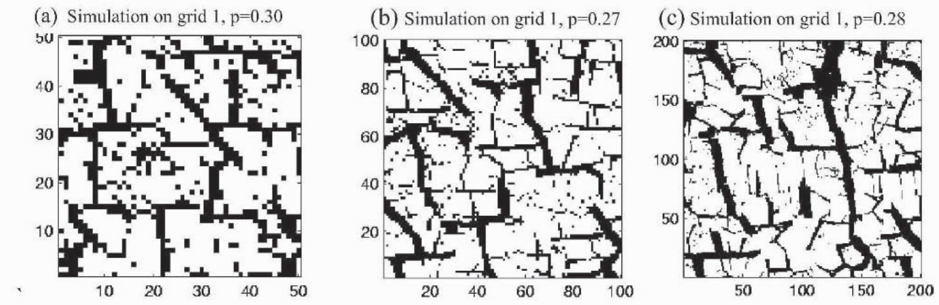


Figure 5. (a)-(c) The same conditional simulation at different scales

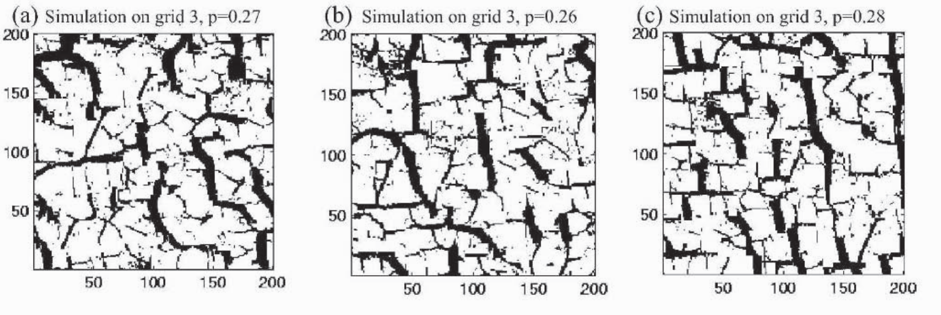


Figure 6. (a)-(c) Three additional conditional simulations at the final grid

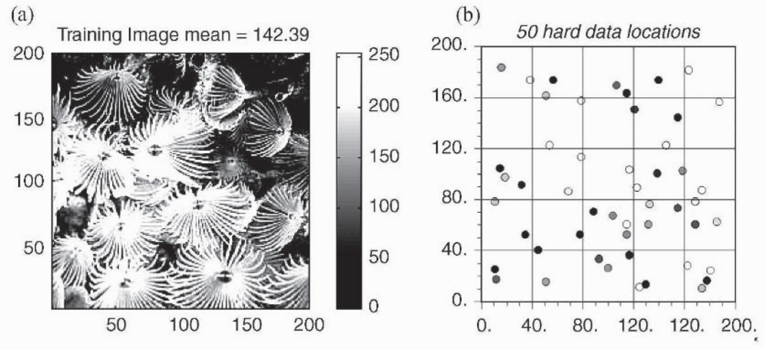


Figure 7. (a) Texture training image, (b) 50 data location map
The same grey scale is used for all maps of Fig.7-8

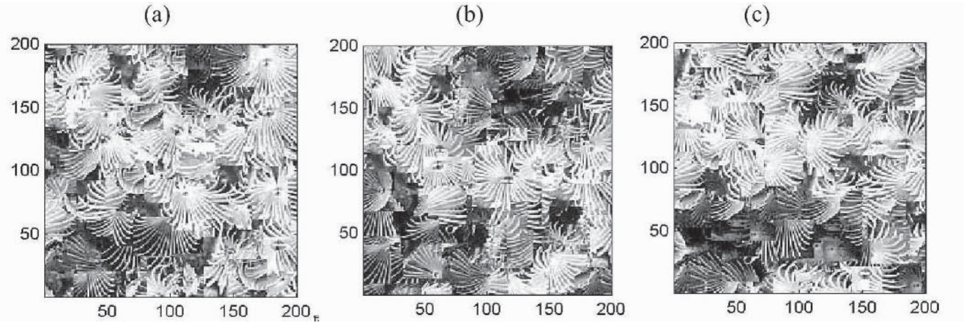


Figure 8. (a)-(c) 3 conditional simulations at the final grid

A continuous variable training image is displayed in Figure 7a, it is a picture of sea anemones. It contains visible gray scale textures with curvatures. Figure 7b shows 50 hard data locations. These samples were generated by sampling a non-conditional simulation from this training image; we used 27x27 pixel templates to classify local training patterns over 3 grid scales. The patch size is 19x19; Figures 8a-8c show three conditional simulations based on the training image of Figure 7a. The simulated anemones can be recognized as such, however with square discontinuities corresponding to the template size. This is the price to pay for working with patches instead of points. This problem calls for future tuning of the algorithm.

It is important to specify correctly both the template size and the patch size. The template size depends on the complexity and the scale of training patterns. The guiding rule is that the template should be large enough that on the coarsest grid it can capture the pattern objects and their interaction. For example, for the sea anemones training image, the size of the largest template should be at least equal to the size of the average anemone object. The patch size can be up to 2/3 of the template size in each direction. A larger patch speeds up the simulation and improves the pattern reproduction, but at the cost of generating discontinuities.

It is suggested to test different template and patch sizes using the full training image but simulating only part of the required field.

6 Conclusions

In this paper, we apply a set of filters to scan training images. Local patterns and textures in training images are classified by a set of filter scores. This leads to a significant dimension reduction of the space of training patterns. Drawing from classes of training patterns allows us to simulate whole patterns as opposed to point values. The simulation proceeds by sequentially visiting each simulation node and identifying the closest training pattern class to the local template data centered at the simulation node. We sample a specific pattern from the identified pattern class, and patch the sampled pattern at the simulation node. Freezing the inner part of patched pattern not only makes the simulation faster but also ensures better pattern reproduction. Multi-grid simulation is implemented, allowing for pattern reproduction at different scales. Although all illustrations are given in 2D with six filters, the algorithm can be extended to 3D using correspondingly 9 filters.

References

Guardiano, F., and Srivastava, R. M., 1993, Multivariate geostatistics: Beyond bivariate moments, *in* Soarses A., ed. Geostatistics-Troia Vol. 1: Kluwer Academic, Dordrecht, pp.133-144

Liang, L., Liu, C., Guo, B., and Shum, H.-Y., 2001, Real time texture synthesis by patch-based sampling, ACM Transactions on Graphics, vol.20, Issue 3, p.127-150

Strebelle, S., 2002, Conditional simulation of complex geological structures using multiple-point statistics: *Math. Geol.,* v.34, no1, p.1-21.

Tran, T. 1994, Improving variogram reproduction on dense simulation grids, *Computers & Geosciences*, vol.25, no.7, pp.1161-68

This page intentionally blank

A PARALLEL SCHEME FOR MULTI-SCALE DATA INTEGRATION

OMER INANC TUREYEN and JEF CAERS
Stanford University, Department of Petroleum Engineering, Stanford, CA 94305-2220, USA

Abstract. In this paper, we propose a parallel modeling approach for solving large and complex inverse problems involving multiple data sources each with different scale of observation. This parallel approach relies on building property models on multiple grids with different resolution at the same time, rather than selecting a single modeling grid. By keeping a high resolution model and its upscaled, coarsened model in constant consistency with each other during the inversion process, a fully consistent integration of all data sources is achieved.

1 Introduction

With the advance of CPU power, numerical models have become an essential part of most engineering applications, be it a finite difference code of flow in the subsurface or a boundary element code modeling the geo-mechanical behavior of faulting and folding structures. Any model, analytical or simulated using a finite element/difference code, is as good as the input material properties on which the physical model is applied. In an Earth Science context, the modeling of such properties is subject to a large degree of uncertainty due to lack of exhaustive access and due to often strong heterogeneity of the medium under study. Instead, a wide variety of indirect data is available to construct various realizations of the media in question. Moreover the various data sources have a different "area of coverage" and "scale of observation". Fine-scale data, for example obtained by drilling a well or by taking a sample at the surface provide direct measurements but are typically sparse in coverage. Remote sensing methods cover a large area but provide only indirect evidence of the properties to be modeled. For the purpose of this paper, the data is subdivided into two parts:

- Static data: refers to direct or indirect observations of the material or rock properties being modeled as input to the physical model. For example a rock/soil type observed in a well, a 3D seismic survey.
- Dynamic data: refers to all data that are direct observations of the physical phenomenon being studied. For example a measurement of pressure/head in a well, a stress or strain measurement.

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 275-284.

The goal pursued in this paper is a method for building property models that honor these two types of data. To address this problem, some difficult challenges will need to be addressed

- The relationship between the dynamic data and the modeled property is in general non-linear, often provided through partial differential equations simulated using a finite element/difference code.
- Static data is often of smaller scale than dynamic data, which provides integrated or convoluted information about the modeled property.

2 Solution using a parallel modeling approach

The purpose of numerical modeling is to predict a response based on a numerical model, e.g. the degree of fracturing in a structure, the production of water or oil in a well. Physical laws on which such prediction rely are generally of the form:

$$f(\mathbf{q}, \mathbf{z}) = 0 \tag{1}$$

where $\mathbf{q}$ are the physical quantities (e.g. pressure, stress), as function of space and time and $\mathbf{z}$ are the material properties, which, in this paper is only a function of space (e.g. porosity, Poisson's ratio). In most cases a 2D or 3D regular or irregular grid of properties $\mathbf{z}$ is generated. The physical law Eq.(1) defines a forward model between the material properties and physical quantities

$$\mathbf{q}^{res} = g(\mathbf{z}, \mathbf{q}^{in})$$

for some initial state and boundary conditions, $\mathbf{q}^{in}$. Static data consist of direct or indirect information related to $\mathbf{z}$, while dynamic data consists of information $\mathbf{q}^{res}$ (e.g. pressure in a well). The non-linear nature of g forces the modeler to use iterative methods for solving the inverse problem (finding $\mathbf{z}$, for given $\mathbf{q}^{res}$), hence requires multiple evaluations of g. When g is a numerical simulation model this may be CPU demanding (Caers, 2004). In that regard the problem of modeling $\mathbf{z}$ calls for a decision on the resolution (dimension) of the modeling grid that is a trade-off between two constraints:

- The grid size should be small enough to include the static fine-scale information, particularly any direct property data.
- The grid size should be coarse enough for finite element/difference simulation of the forward model g to be feasible within reasonable CPU-time. This is important for including the dynamic data on $\mathbf{q}$

In this paper, we propose a parallel modeling approach for $\mathbf{z}$ that avoids this trade-off by working on two grids at the same time: a high resolution grid that allows including any fine-scale static information on $\mathbf{z}$ and a coarsened grid that allows running multiple evaluations of g and thereby including the dynamic data on $\mathbf{q}$. The key idea presented in this paper is to keep the two grids in constant consistency with each other both in terms of the property $\mathbf{z}$ and the responses on $\mathbf{q}$.

The proposed parallel modeling approach follows the following basic steps.

Step 1: High resolution model generation

A high resolution geostatistical realization ($\mathbf{z}$) that honors the static data is generated. This initial high resolution model does not yet match the dynamic data, hence will need to be perturbed. Any perturbation method can be used (gradient-based, Metropolis samplers, rejection methods etc,...) In this paper, such perturbations are represented by a set of parameters $\mathbf{r}$ that change a realization $\mathbf{z}$ into a perturbed realization $\mathbf{z}(\mathbf{r})$, the magnitude of perturbation given by $\mathbf{r}$. When $\mathbf{r}$=$\mathbf{0}$, no perturbation is performed hence $\mathbf{z}$=$\mathbf{z}(\mathbf{0})$.

Step 2: Optimized gridding and upscaling

The next step consists of upgridding and upscaling the high resolution realization $\mathbf{z}(\mathbf{r})$ to a coarsened realization $\mathbf{z}^{up}(\mathbf{r})$: relationship:

$$\mathbf{z}^{up}(\mathbf{r}) = S_\theta(\mathbf{z}(\mathbf{r})) \tag{2}$$

Here S_θ represents the upscaling/upgridding technique applied on $\mathbf{z}(\mathbf{r})$ and θ the set of upgridding parameters (grid dimensions, averaging type, etc..) Upgridding refers to the construction of the coarse grid, which could be Cartesian or irregularly gridded. The dimension of the coarse grid is defined by the number of grid vertices in each x, y and z-direction. Upscaling refers to methods for assigning coarse grid properties given the high resolution property realization. Due to the presence of upscaling/upgridding errors, the high resolution and coarsened realization may conflict in terms of the responses when the forward model is applied on each of them. This possible inconsistency between the two grids needs to be reduced by minimizing the upscaling/upgridding errors introduced. Define as "true" but unknown upscaling error, the difference between responses of the high resolution and coarsened model:

$$\epsilon = \|g(\mathbf{z}^{up}(\mathbf{r})) - g(\mathbf{z}(\mathbf{r}))\| \tag{3}$$

This error cannot be calculated since the forward model g is too CPU-demanding to be evaluated on the high resolution realization. Hence the challenge is to reduce the upscaling error ϵ without knowing $g(\mathbf{z}(\mathbf{r}))$. To achieve this, we introduce a function g^* as an approximation to g that is less CPU demanding to evaluate. The shape of g^* is problem specific and could be a model with simplified physics or could be an analytical model (see example section for specifics). The approximate forward model allows to approximate the true upscaling error

$$\epsilon^*(S_\theta) = \|g^*(S_\theta(\mathbf{z}(\mathbf{r}))) - g^*(\mathbf{z}(\mathbf{r}))\| \tag{4}$$

The key idea is to reduce ϵ by reducing ϵ^*. ϵ^* can be reduced by adjusting the parameters θ of the upscaling/upgridding method until Eq.(4) is minimized. At the same time Eq.(3) will be minimized if the ranking of models $\mathbf{z}$ provided by the forward model g^* is the same as the ranking provided by g. In other words, ϵ and

ϵ^* need not be the same in absolute magnitude, ϵ^* must decrease when ϵ decreases and vice versa.

Step 3: Mismatch calculation and perturbation

Once an optimized coarsening of the high resolution model is determined, a model response is obtained by evaluating the forward model (g) on the coarsened realization

$$\mathbf{RP}_{\mathbf{Z}^{up}}(\mathbf{r}) = g(\mathbf{z}^{up}(\mathbf{r}))$$

the $\mathbf{r}$ parameters can be optimized by minimizing the difference between the coarsened model response $\mathbf{RP}_{\mathbf{Z}^{up}}(\mathbf{r})$ and the dynamic data $\mathbf{D}$:

$$\min_{\mathbf{r}} O(\mathbf{r}) = \min \parallel \mathbf{RP}_{\mathbf{Z}^{up}}(\mathbf{r}) - \mathbf{D} \parallel$$

Some important properties of this approach are

- All data, fine-scale static and coarse-scale dynamic are integrated simultaneously.
- The upscaling/upgridding optimization forces the high resolution and coarsened model to be consistent with each other during the complete inversion process.
- The properties are modified on the high resolution grid, hence any model perturbation $\mathbf{z}(\mathbf{r})$ can be kept consistent with the static data and prior geological information.

3 Parallel modeling in reservoir characterization

3.1 RESERVOIR MODELING

In this section we present in greater detail how the parallel modelling methodology is applied to the problem of reservoir characterization for hydrocarbon reservoirs along with example applications.

Reservoir modelling calls for the integration of various data into a single reservoir model. Such data sources can be classified as follows:

1. Geological interpretation of reservoir architecture at all scales ranging from major faults to facies and bedding configurations. In geostatistics, such information can be quantified through the variogram or through 3D training images.
2. Well-log and core measurement information is often the most direct type of information, however is only telling of the near well-bore reservoir heterogeneity and provides information at a small (cm) scale.
3. 3D seismic information is probably most exhaustive, yet often at a scale larger than the reservoir modelling scale. In geostatistics this type of data is treated as soft, static data.

4. Reservoir dynamic data, most particularly from pressure and flow measurements, or increasingly common, 4D seismic. The scale of information provided by this kind of data set is largely unknown. It is spatially varying and dependent on boundary conditions and configurations of wells.

The current practice of reservoir modelling consists of modelling the reservoir first using the static data (sources one to three) on the high resolution grid, then upscaling the high resolution model into a coarsened model on which flow simulation (function g) is feasible. Next, the coarsened model is further adjusted to match the dynamic information (source four). In reservoir engineering this is commonly known as "history matching".

Such an approach has the following drawbacks:

- Any high resolution reservoir information (core or well-log) may be lost when the coarsened model is changed.
- Important fine and coarse scale geological information may be destroyed while history matching. Particularly when the history matching method does not take into account statistics such as variogram or multiple-point statistics that are characteristics particular to the high resolution geological model.

3.2 METHODOLOGY OVERVIEW

We provide first a broad overview of a parallel modelling scheme specific to reservoir characterization see (Figure 1).

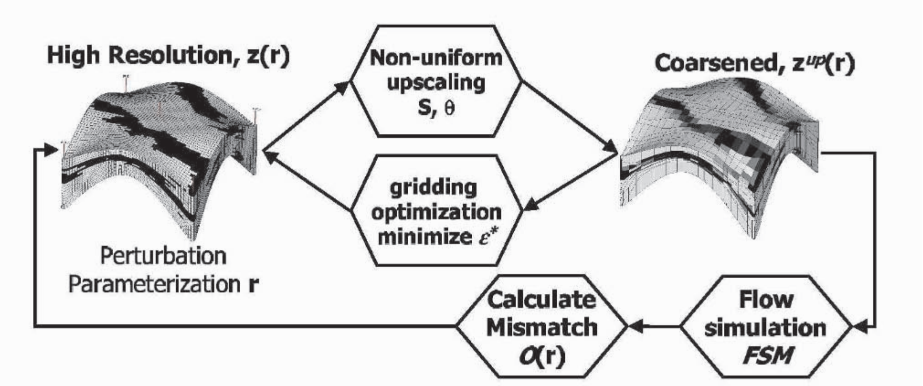

High Resolution, z(r)
Non-uniform upscaling S, θ
Coarsened, z^{up}(r)
gridding optimization minimize ε*
Perturbation Parameterization r
Calculate Mismatch O(r)
Flow simulation FSM

Figure 1. Parallel approach for reservoir characterization

The work-flow starts by constructing a high resolution realization ($\mathbf{z}(\mathbf{r})$) which can be perturbed using a set of perturbation parameters $\mathbf{r}$. A gridding optimization is performed in order to obtain the "non-uniformly gridded" coarse model which minimizes the mismatch between the flow responses of the high resolution model and the coarsened model. The optimization is performed on the θ vector (shown in Figure 1), which represent the various gridding parameters specific to the gridding algorithm (see next section). Once the optimally gridded coarse model is obtained, flow simulation (denoted by *FSM* in Figure 1) is performed on the coarsened model. Then the mismatch between the observed field data and the simulation results are compared. The **r** parameters are adjusted to reduce the mismatch. The entire loop is repeated until this mismatch is minimized.

3.3 PERTURBATION PARAMETERIZATION

Various perturbation methods can be used to perturb the high resolution geological model. However, perturbations should be parameterized such that any perturbation $\mathbf{z}(\mathbf{r})$ honors the same geological continuity model (variogram, object model, training image model) and the same static data as the initial high resolution model $\mathbf{z}$. The gradual deformation (Hu and Roggero, 1997) and the probability perturbation method (Caers, 2003) methods are two examples of perturbation that honor this information.

3.4 GRIDDING - 3DDEGA

Although the above outlined parallel modeling approach is general in the type of gridding method, in this section we review an existing gridding algorithm, 3D-DEGA (3D Discrete Elastic Grid Adjustment, (Garcia, Journel and Aziz, 1992)). The algorithm is devoted to the generation of quadrilateral or hexahedric grids suitable for grid adaptation based on reservoir properties (ϕ, k), pressure fields, saturation fields or any other variable. The resulting grids are in a corner point geometry fashion and can be used with most commercial flow simulators. The main idea behind the 3D-DEGA algorithm is to generate coarse grid blocks that are as homogeneous in terms of a given input variable or variables (permeability, porosity, facies map, etc.).

3.5 GRIDDING OPTIMIZATION, APPROXIMATE FORWARD MODEL

The forward model g is a flow simulation that includes all physics necessary (gravity, capilary pressure, compressibility, multiple phases etc..) for simulating the actual reservoir flow. In real cases it may take several hours to run a full flow simulation on a grid of the order of 10^5 cells depending on the complexity of the physical model. To incorporate static information from well-logs and seismic, a typical high resolution geostatistical realization generated using stochastic simulation can be of the order of $10^6 - 10^7$ cells, a resolution on which flow simulation is not feasible. Approximate flow simulations are therefore required to minimize any upscaling errors as outlined above.

As an approximate physical model we use an incompressible single phase flow simulation (denoted by g^*). A single phase flow model can be calculated on the high resolution grid in a matter of minutes. To further aid the upgridding method we trace streamlines using the pressure (and velocity) solution of the single phase flow model. Using the streamlines simulation, approximate flow responses can be obtained, denoted by $\boldsymbol{RP}^*$.

The high resolution geostatistical model is then upscaled/upgridded to a coarser model with 3D-DEGA given some gridding parameters θ. The incompressible single phase flow solution and streamline simulations are repeated on the coarsened model, from which the same type of responses as for the high resolution model are calculated. The responses from the high resolution and the coarsened model are then compared. The gridding parameters θ are adjusted in an iterative manner until the mismatch (Eq.(4)) between the two responses is minimized.

3.6 2D EXAMPLE

3.6.1 *Definition of the problem*

A 2D, high resolution reference permeability realization is generated through a training image based technique using multiple point geostatistics. This was performed using the SNESIM (Strebelle, 2002) algorithm with the training image given in Figure 2a. The realization is representative of a channel system with a quarter five spot pattern production strategy where injector producer wells are placed on opposite corners. Water flooding for 500 days is simulated on this realization using a finite difference reservoir simulator (ECLISPE). A water cut curve is obtained (see reference permeability map and its corresponding water cut curve in Figure 2b and 2c). This curve is treated as dynamic data. A constant permeability value of 10000 md is assigned to the sand (channel) facies and 100 md is assigned to the mud (non-channel) facies. The permeability values at the well locations are assumed "known" and are treated as "hard data". This high resolution realization is composed of 100 × 100 × 1 grid blocks in the x, y and the z directions, where each block is of size 20ft × 20ft × 200ft.

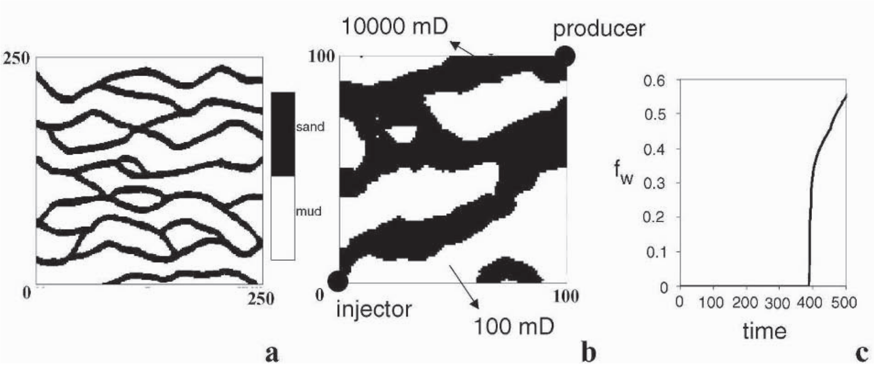


Figure 2. (a) The training image used for creating the reference permeability field, (b) The reference permeability field, (c) The flow response of the reference permeability field.

In order to demonstrate the effectiveness of the parallel modeling approach, 30 high resolution realizations were generated (using SNESIM), conditioned only to hard data (no history matching was performed). Full flow simulation was performed on these 30 realizations and Figure 3 illustrates the flow responses. As expected the flow responses of the high resolution models conditioned only to hard data provide a wide scatter of production responses.

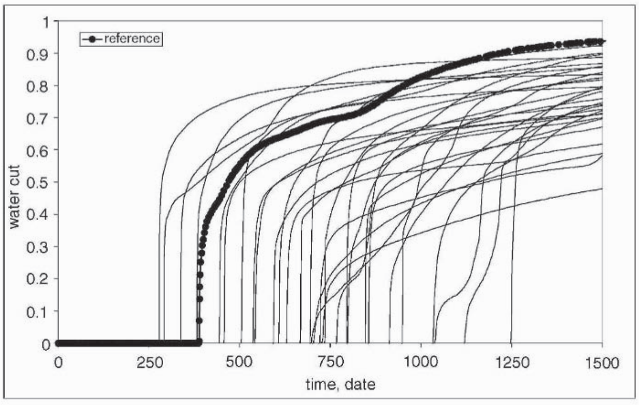


Figure 3. Flow responses of 30 realizations not matched to reference water cut data.

3.6.2 *Applying the parallel modelling approach*

Figure 4 illustrates the work-flow of the parallel modeling methodology specific for this example. In the first step the SNESIM algorithm is used for generating a high resolution realization (with the same dimensions of the reference realization). In generating this initial realization, the same training image given in Figure 2a is used. Streamline simulation (which acts as the approximate flow model, denoted by g^* in the previous section) is performed on this high resolution realization and a pseudo (approximate) water cut curve ($\mathbf{RP}^*$, the approximate flow response) is obtained.

Using the 3D-DEGA algorithm (with an initial guess of upgridding parameters θ), the high resolution model is upgridded and upscaling is performed by taking arithmetic averages of the permeability values from the high resolution realization. Streamline simulation is performed on the coarsened model and a pseudo water cut curve ($\mathbf{RP}^{up*}$) is calculated. An optimization step minimizing the mismatch between $\mathbf{RP}$ and $\mathbf{RP}^{up*}$ is applied, during which the upgridding parameters θ are optimized. In addition to the upgridding parameters, the number of coarse grid blocks are also optimized while keeping the total number of grid blocks constant ($n_x \times n_y = 625$).

Full flow simulation (the full flow model, denoted by g in the previous section) is performed on the optimally gridded coarsened model and the simulated water cut curve is obtained. The mismatch between the water cut curve and the reference water cut curve (given in Figure 2c) is computed. The perturbation method (probability perturbation, (Caers, 2003)) repeats the entire procedure until this mismatch is minimized by optimizing the $\mathbf{r}$ parameters.

3.6.3 *Results for the 2D synthetic example*

The proposed parallel modeling approach has been performed initially without the gridding optimization for 30 different random seeds in order to make a comparison with Figure 3. At the end of the inversion process, for each random seed, a coars-

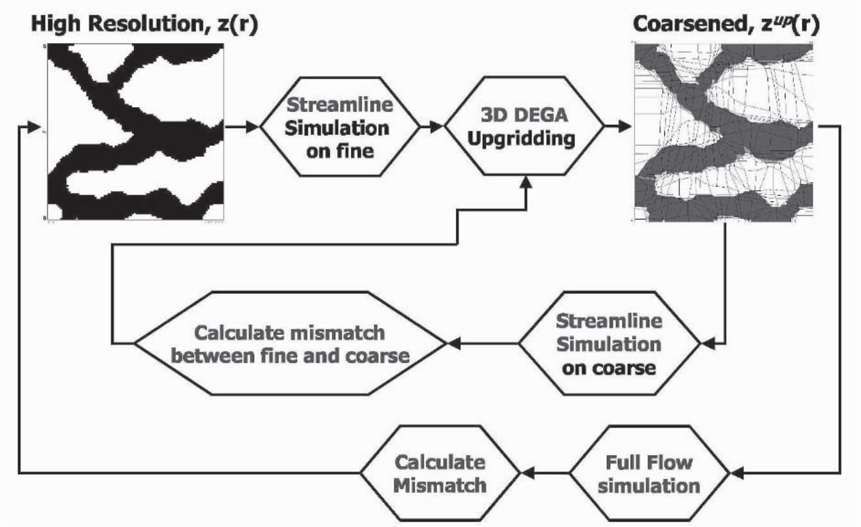

High Resolution, z(r)
Coarsened, $z^{up}(r)$
Streamline Simulation on fine
3D DEGA Upgridding
Calculate mismatch between fine and coarse
Streamline Simulation on coarse
Calculate Mismatch
Full Flow simulation

Figure 4. Schematics of the parallel approach specific to the 2D example.

ened and a high resolution flow response results as output. The coarsened models match the history well, and provide accurate future predictions for the same well configuration (see Figure 5a). Figure 5b illustrates the flow responses of the high resolution models corresponding to the history matched, coarsened models. As it is clear from the Figure 5b, the high resolution models in this case provide a match to some degree when compared with Figure 3, but not as well as the match of the coarsened models. This can be attributed to upscaling errors that introduce inconsistency between the two modeling solutions.

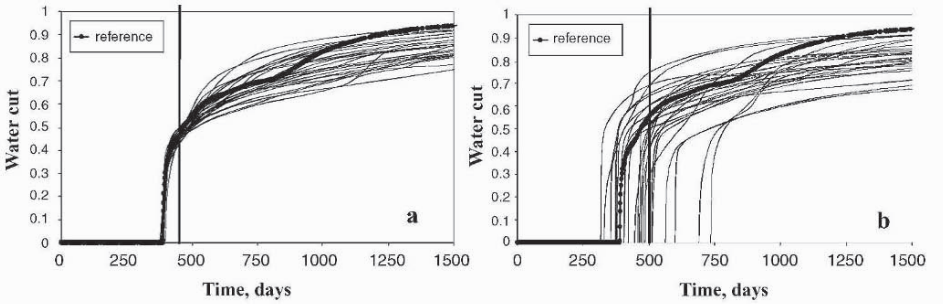


Figure 5. Flow results of the parallel modelling approach without the gridding optimization. (a) Flow responses of the non-optimally gridded coarse models, (b) Flow responses of the corresponding high resolution models.

Therefore, gridding optimization needs to be introduced each time a high resolution model coarsened. Results obtained by incorporating the gridding optimization shown in Figure 6a for the coarsened model match the history well. Figure 6b illustrates the flow responses of the corresponding high resolution models. Improvements on these responses are clear when compared with Figure 5b.

The scatter on the high resolution model response is reduced considerably through gridding optimization.

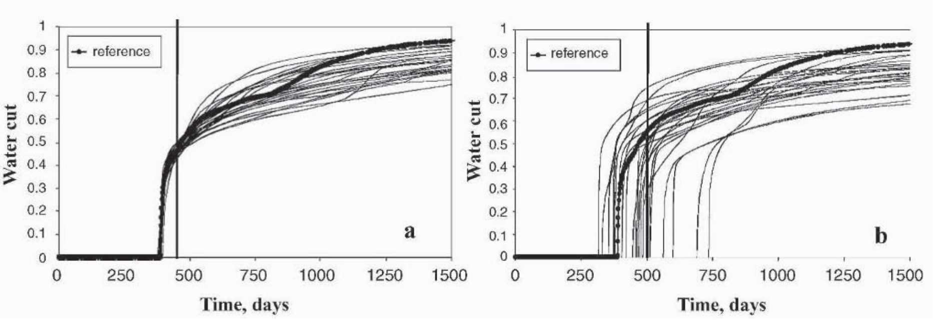


Figure 6. Flow results of the parallel modelling approach. (a) Flow responses of the optimally gridded coarse models, (b) Flow responses of the corresponding high resolution models.

4 Conclusions

In most earth sciences applications "modeling" in general is a difficult task due to lack of exhaustive data and heterogeneity of the medium under study. Fully integrating all data into a single grid is usually not plausible.

The parallel modeling approach uses multiple modeling resolutions at the same time. A (uniformly gridded) high resolution model is used for integrating static data and applying geostatistics effectively. The coarsened model may be gridded in any fashion and is used for integrating dynamic data. Consistency between these two resolutions of models is ensured through a gridding/upscaling optimization step, which forces the responses of each model resolution to be consistent.

References

Garcia, M. H. Journel, A. G. and Aziz, K. *Automatic Grid Generation for Modeling Reservoir Heterogeneities*, SPE Reservoir Engineering, May, 1992, p. 278.

Caers, J. *Geostatistical History Matching Under Training-Image Based Geological Model Constraints*, SPE Journal, September, 2003, p. 218-226.

Roggero, F. and Hu, L. Y. *Gradual deformation of continuous geostatistical models for history matching*, SPE 49004, presented at the SPE Annual Technical Conference, 27-30 September 1998, NewOrleans-Lousianna-USA.

Strebelle, B.S. *Conditional Simulation of Complex Geological Structures using Multiple-Point Geostatistics*, Math. Geol., v34, no.1, January, 2002, p1-22.

Landa, J.L., Horne, R.N. *A Procedure to Integrate Well Test Data, Reservoir Performance History and 4-D Seismic Information into a Reservoir Description*, SPE 38653, presented at the SPE Annual Technical Conference, 5-8 October 1997, San Antonio-Texas-USA.

Caers, J. *Data Conditioning With the Probability Perturbation Method*, Paper presented at the 7^{th}International Geostatistics Congress, 26 September - 1 October 2004, Banff, CANADA.

STOCHASTIC MODELING OF NATURAL FRACTURED MEDIA: A REVIEW

JEAN-PAUL CHILÈS
Centre de géostatistique, École des mines de Paris,
35 rue Saint Honoré, 77305 Fontainebleau cedex, France

Abstract. The connectivity and the flow or mechanical properties of networks of faults and joints are key factors in a number of applications. Only a minute fraction of the fractures in the domain of interest are usually observed, so that a deterministic modeling of the fracture network is not possible. Stochastic models have been developed for a variety of fracture patterns. They can be classified in objects models, which are purely stochastic, and process-based models, which take account of the mechanical processes that rule fracturing. This presentation is focused on the geometrical and topological aspects of fracture networks.

1 Introduction

Structural discontinuities such as faults and joints occur at various scales and are now widely recognized as a key factor in a number of situations. They can act as conduits or flow barriers depending on whether they are open or sealed, and thus impact the safety of nuclear waste storage in geological formations, the oil recovery in fractured hydrocarbon reservoirs, and the heat recovery in geothermal hot dry rock reservoirs. Open fractures delimit blocks of ornamental stones or impact the stability of stopes or the safety of underground exploitations, caverns and tunnels. Veins are associated with mineralization.

Fracturing results from a number of processes and depends on the rock fabric, rock properties, geological setting, past and present mechanical constraints. Many different fracture patterns can therefore be observed and there is no general fracture network model. While large faults are usually known from geological mapping and seismic surveys, medium and small scale structures are very sparsely observed and are therefore represented by means of stochastic models. Many purely stochastic models, where the fractures are represented by objects, have been developed. The mechanical processes leading to the initiation and growth of faults and joints are better and better understood, which has lead to the development of algorithms modeling the fracturing process or at least parameters that control it. These algorithms are exploited by process-based models. We will first review the main types of object models and then examine approaches based on a modeling of geological and mechanical processes. This paper will be focused on the geometrical and topological aspects of fracture networks, but in applications this shall not be disconnected from the hydrological or mechanical aspects.

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 285-294.

2 Object Models

2.1 BASIC RANDOM SET MODELS AND GENERALIZATIONS

The first fracture network models were deterministic, such as the model defined by three orthogonal series of parallel and regularly spaced planes. Such a network is too regular, subdividing the space like sugar cubes, so that stochastic models were soon proposed. The first stochastic models were simple prototypes corresponding to the usual models of random set theory. Priest and Hudson (1976) used random planes to represent fractures that can be considered as infinite at the scale of the domain under study. To represent finite-size fractures, Baecher et al. (1977) developed the random disk model, namely a standard Boolean model where the objects are disks with random diameters and orientations. Another way to obtain finite fractures is to start from a random plane model and define a stochastic tessellation in each plane, for example a Voronoi or Poisson polygons tessellation; each polygon is then randomly considered as a fracture or as intact rock. Depending on the method chosen for defining the tessellation, the fractures can intersect with others (Veneziano, 1978) or abut against others (Dershowitz, 1984).

These basic models represent networks with a uniform fracture intensity. In applications, however, fracture intensity is usually not uniform. This is accounted for by generalizing the Boolean model, based on Poisson points, to marked point processes based on more general point processes. The main point process models used are (e.g., Stoyan and Stoyan, 1994): (i) Poisson point process with spatial variations of intensity, these variations being either deterministic (inhomogeneous Poisson point process) or modeled as a positive stationary random function (Cox process); (ii) cluster process (also called shooting process, or parent-daughter model): primary points (targets, or parents) constitute a Poisson point process; each primary point is the center of a cluster of secondary points (shot impacts, or daughters) randomly and independently located around the primary point according to some dispersion distribution; (iii) hard-core model, to forbid the presence of too close points, for example because the relaxation of constraints in the vicinity of a fracture inhibits the creation of a new fracture. Chilès (1989) modeled fractures in a granitic site as clusters at a local scale with a regionalized intensity at a wide scale.

In practice, the orientation of the fractures is not purely random, and several fracture sets are superimposed, each one with a direction that is well defined or varies around a mean direction. Each set is linked to an event of the structural history of the site and has its own characteristics as regards to the size of the fractures, their aperture, etc.

2.2 HIERARCHICAL MODELS

Most of the models presented above were developed in view of mining engineering applications or for the study of potential nuclear-waste underground storage sites in granitic formations, namely for situations where there is no well-expressed hierarchy among the various fracture sets, despite of their chronology. The situation is very different for sedimentary rocks, so that other models were developed for them.

Hierarchical models are more complex than basic models because they must include the relationships between the various fracture sets. Conversely, fractures in sedimentary rocks are often either horizontal (bedding planes) or subvertical and confined to one or several layer (joints), so that their 3D modeling amounts to a series of 2D models. The first stochastic hierarchical model we know is a 2D model in a petroleum reservoir (Conrad and Jacquin, 1973): (i) large faults are represented by a primary network of random lines; (ii) in each Poisson polygon defined by the primary network, an independent Boolean model of segments represent finite fractures; the model is truncated by the boundaries of the polygon, so that fractures can abut against faults of the primary network but cannot intersect them.

Bourgine et al. (1995) propose another model, developed to model vertical joints at a pluridecametric scale in the Saq sandstones, which are considered as an analog to some petroleum reservoirs. That model is based on renewal processes, which offers much flexibility: (i) a primary set is composed of subparallel finite fractures; (ii) secondary sets connect fractures of the primary set, with a given proportion of terminations abutting against fractures of the primary set. A 3D model is obtained by superimposing several such models with parameters depending on the bed height. Other models can be found in the literature.

2.3 MULTISCALE BEHAVIOR AND FRACTAL MODELS

The organization of fractures, more precisely of faults, is often considered as self-similar (e.g., Turcotte, 1992) because faults can occur at all scales from cartographic faults to microfaults. This is often sustained by the fact that some variable (e.g., fracture size, fracture spacing, size of rock fragments, box counting, i.e., number of cells intersected by fractures as a function of cell size) follows a power law distribution with a noninteger parameter. In conclusion of a study of several sites in sandstones, Gillespie et al. (1993) observe that the spacings between tectonic faults follow a power law distribution because of a clustering of smaller faults in the vicinity of larger faults, whereas the spatial distribution of major joints is very regular because it is controlled by the bed thickness of the jointed unit, by the differences in mechanical properties between the jointed unit and adjacent layers and by the extensional strain. The situation is different in granitic rocks where the spacing between joints can follow a power law distribution (Barton and Zoback, 1990), whereas joint corridors and faults can be regularly spaced (Genter and Castaing, 1997). So the invariance laws (self-similarity or self-affinity ruled by a power law) which are often advocated have no universal character and shall be used only between well identified bounds, as noticed by Hatton et al. (1994) and Peacock (1996). Moreover, the observation of a power law distribution for some parameter of the fractures does not imply the fractal character of the fracture network, as shown for example by Walmann (1998) in a laboratory experiment where clayey material was submitted to mechanical tests generating cracks.

In comparison with the abundant literature about fractals, few fractal models have been proposed to represent fracture networks, probably because one or several fractal exponents are far from characterizing a fractal model. Let us mention models defined by fragmentation (Turcotte, 1986; Acuna and Yortsos, 1995): the domain of interest is subdivided in two parts by a fracture of the first generation; each part in then subdivided in two parts by a fracture of the second generation, and so on. To obtain a fractal model, the fractures are not created systematically but with a given probability; when a fracture is not created the subblock is no more subdivided. Bour and Davy (1999) developed another model, which is the transposition of the random disk model to fractals, by locating disks with a power law diameter distribution at the points of a fractal point process.

These synthetic models are often used to study the connectivity or flow behavior of the network—percolation, emergence of a continuous-medium behavior—as a function of the fractal dimension and other fractal exponents, either analytically or by simulation. The validity of the results for other models with the same fractal exponents, often assumed, is questionable.

The detailed fracture data sets studied by this author did not show evidence of a fractal behavior, either locally (Chilès, 1988, for granite) or over a wide range of scales (Castaing et al., 1996, for sandstones). In the latter case, the analysis showed a very different behavior for faults and joints, as well as characteristic scales for the joints, which could be related to the various mechanical units formed by the sedimentary beds, the sandstone formation, the sedimentary basin, and the upper crust. In such a case, stochastic models are built at a given scale, with the fractures of the finer scale incorporated with the rock matrix and the fractures at the coarser scale—usually few in number—modeled deterministically. Other models are needed for networks that are not controlled by mechanical units, for example in the sandstones studied by Odling (1997).

2.4 STATISTICAL INFERENCE

The key problem, from a geostatistical perspective, is the inference of the parameters of the stochastic models. Even for a simple model like the random disk model, this is not trivial, because fractures are 3D objects that cannot be observed directly but only through their intersections with boreholes (cores, electric logs) or outcrops (field exposures, drift walls, vertical stopes, aerial photographs). Part of the problem is also that the fractures of our models are an idealization of reality: true fractures are not planar circular disks for example. So if the choice of the model is not sound, it may be difficult for it to honor a variety of statistics about fracture density, fracture orientation, trace length, abutting fractures, connectivity, etc.

The most significant parameter is the fracture density μ, defined as the average fracture surface per volume unit. If we consider a set of fractures which are normal to a sampling line, this is simply the number of fractures per length unit; this is also the inverse of the average fracture spacing. Fractures that are oblique to a surveyed line or outcrop are underrepresented in comparison with fractures orthogonal to it (fractures parallel to the line or outcrop cannot be observed). It is therefore recommended to have several sampling lines or outcrops with different orientations, and data must be weighted according to the orientation of the fracture with respect to the line or outcrop: this is the aim of the well-known correction proposed by Terzaghi (1965) for line sampling, which can be improved in several ways (Yow, 1987) and has a variant for areal sampling (Chilès and de Marsily, 1993).

The fracture density μ is usually not the basic parameter of a fracture network model. For a random disk model for example, with disks of a fixed orientation, the parameters are the Poisson process intensity λ, namely the number of disks per volume unit, and the distribution of disk diameters F_D. It could be tempting to separately infer both parameters, as Zhang and Einstein (2000), for example, do it. These parameters, however, are not robust: it is not obvious to decide on the field that two aligned traces separated by a short intact interval are two distinct fractures or the en-échelons part of a single fracture, and structural geologists can adopt either view. The choice, however, will not affect the calculation of the total fracture length and thus the estimation of the fracture density μ, which is therefore a robust parameter. Consequently, it is preferable to infer separately the fracture density μ and the disk diameter distribution F_D, and then derive the corresponding Poisson process intensity λ by applying the relation $\mu = \pi \lambda M_2 / 4$, where M_2 is the quadratic mean of the disk diameter.

Now the diameters of the fractures cannot be measured because fractures are at best observed as fracture traces. The length of a fracture trace is shorter than the diameter of the fracture but the disk diameter distribution can be derived from the trace length distribution using the stereological formula which is appropriate for the kind of sampling used (e.g., Lyman, 2003). The determination of the trace length distribution, however, requires special care to take the various sources of bias into account: overrepresentation of long traces, truncation of the short traces, censoring when the terminations cannot be seen, etc. (see Lantuéjoul et al., 2004, and references therein).

Things are not simpler for the inference of the parameters of more complex models. For those deriving from the Boolean model, tools specific to point processes, such as Ripley's *K*-function and its variants for isotropic point processes and the pair correlation function can be applied to fracture centers (Stoyan and Stoyan, 1994; Wen and Sinding-Larsen, 1997). In many cases, however, the fracture trace centers are not precisely known due to censoring (at least one termination of the trace cannot be observed), which limits the use of these tools. With outcrop data for example, it could be more appropriate to use tools specifically designed for random fiber processes (Schwandtke, 1988). The spatial variability can also be studied with the variogram of the observed fracture density, defined as the cumulated fracture length in equal squares partitioning an outcrop, or as the number of fracture intersections in equal segments partitioning a borehole.

All these tools as well as simple statistical tools can be used qualitatively to assess the choice of a relevant model. For example, the distribution of the spacing between successive fractures of a given set is exponential if the fractures are randomly located, less dispersed in the case of a regular pattern, or more dispersed, with many short spacings and a long tail, if the fractures are clustered. To transform them into quantitative tools, it is necessary to know their theoretical expression as a function of the parameters of the model; these expressions must incorporate the bias sources and stereological relationships. General results are given by Pohlmann et al. (1981). A suitable approach has been developed for as complex models as the disk cluster model with regionalized intensity (Chilès, 1989) and the hierarchical model proposed by Bourgine et al. (1995).

3 Process-Based Models

3.1 IDENTIFICATION AND MODELING OF MECHANICAL PROCESSES

The models presented above place fractures in their final state. Another approach consists in modeling the fracturing process itself. That approach is theoretically more satisfactory and also allows the prediction of the future evolution of the system in response to a new tectonic event or underground works. The identification of fracturing processes motivated a detailed observation of natural fracture systems: fault zones in sandstones (Antonellini and Aydin, 1994, 1995) and granitic rock (Christiansen and Pollard, 1997), joint formation in granitic rock (Segall and Pollard, 1983), relation between faults, joints and stress (Finkbeiner et al., 1997). Laboratory experiments were also carried out on scale models to observe the nucleation and development of discontinuities: brittle varnishes for understanding the mechanical origin of joints in layered media (Rives and Petit, 1990), sand and silicones (Sornette et al., 1990) or clay and fault gouge (An, 1998; Walmann, 1998) for studying the origin of faults in the Earth's crust. Finally, numerical codes have been developed to model these mechanical processes, usually under the simplifying assumption of an elastic stress field (e.g., Thomas, 1993). Such codes have been applied to model fault growth, linkage and interaction (Aydin and Schultz, 1990; Bürgmann et al., 1994; Willemse et al., 1997; Crider and Pollard, 1998; Weinberger et al., 1999) and fault-related fracturing (Martel and Boger, 1998). Caputo and Santaroto (1998) developed a mechanical model for quantifying the ratio between extensional joints and faults.

3.2 STOCHASTIC AND PROCESS-BASED APPROACH

The numerical modeling of fracture initiation and growth is not a simple task and is carried out for rather small networks in comparison with the capabilities of flow simulators, which can handle one million fractures when the rock matrix is impervious. This has led to the development of iterative techniques that simulate the initiation of new fractures and their growth. The growth is stochastic rather than the result of a mechanical calculation but the rules that govern the direction and intensity of growth are based on mechanical principles (Takayasu, 1985; Renshaw and Pollard, 1994).

Bourne et al. (2000) are a typical example of that kind of approach. They integrate growth processes in a geomechanical model of rock deformation which is the basis for a simulation of the fracture network. Large discontinuities such as seismically visible faults are supposed to be known. The first step is to determine the stress field within the faulted reservoir; this is done by assuming that the rocks behave as a homogeneous, isotropic, and linear-elastic material, and the faults as surfaces free of shear stress. The distribution of elastic stress related to faulting is governed by the distribution of slip over the fault network. Since fault displacement observations are scarce and poorly reliable, the distribution of slip is calculated over the fault network by loading it according to the remote stress that caused the faults to slip (Jeyakumaran et al., 1992). The orientation of this remote stress is estimated from the regional tectonic history, and its magnitude according to the mean rock strength prior to faulting. The comparison of the elastic stress field with the brittle failure strength of the reservoir rock determines the areas where secondary tensile and shear fractures have occurred. The initiation, growth, and termination of these fractures are then simulated in these areas. During growth, fracture spacing and interaction are controlled: a forbidden zone is defined around each fracture, which represents an overall reduction in local stress due to the presence of the fracture; conversely, the mechanical interaction leading to the connection of fractures with neighboring tips is taken into account, leading to en-echelon structures and enhancing the connectivity of the fractures

All the elements needed by the approach above are not always available. It is however useful to account for geomechanical information. Srivastava (2002) defines a simplified approach to generate 3D simulations of a fracture network observed as lineaments at a regional scale (40 × 50 km) on well-exposed areas of the topographic surface. The fracture traces which have not been observed or whose terminations could not be seen are obtained by simulating their growth. That growth is guided by a statistical model rather than determined by a mechanical process. The statistical model rests on a detailed analysis of the distribution of the characteristics of the fractures (length distribution, dip distribution, etc.) and of their correlation (variograms). The vertical growth is simulated similarly for all fractures according to geomechanical assumptions about their shape.

3.3 INTEGRATION OF AUXILIARY INFORMATION

An intermediate solution between object models and process-based models is to model the fracture network by placement of objects but to determine the local parameters of the object model by using geomechanical rules. That approach is used for example by Cacas et al. (1997) to simulate large joints in stratified sedimentary reservoirs: the local direction of the systematic joints is deduced from a mechanical simulation of the local stress and strain tensors at the time of fracturing; the direction of fold-related joints, which is parallel to the hinge of the fold, is defined locally by a surface curvature analysis; etc.

Similarly, the fracture density is larger in the vicinity of an anticlinal axis, where the layers were submitted to an extensional regime, than on its flanks. Like in fault zones, it is then necessary to build a model of the spatial variations of fracture density. In more complex situations, the strain field can be reconstructed by finite element methods from outcrop measurements (Schultz-Ela, 1988) and its impact on the fracture network can be modeled.

The situation is more complex if the evolution of the structural setting concerns not only the fracture density but also the fracture type. For example, in the sandstones of Arches Park (Utah), joint corridors with a large permeability can be observed in the vicinity of anticlinal axes, whereas deformation bands are found in neighboring synclines, namely structures resulting from a strong compression of grains and thus with a very low permeability (Antonelli and Aydin, 1994, 1995).

4 Conclusion

With the increasing power of computers, stochastic and process-based methods shall be able to simulate realistic fracture networks. However, it will remain difficult to reproduce the exact characteristics of a given site: they depend on several factors, including the spatial variability of rock properties, which is poorly known. Intermediate approaches are therefore useful. They can integrate geomechanical rules, field data, and complementary data such as new seismic attributes brought by high resolution 3D seismic surveys (Gauthier et al., 2003).

Simulated fracture networks are usually the input of flow models. Conversely, the flow behavior of the true network can help in choosing the relevant fracture network pattern and its parameters. Inverse methods should be developed to that effect.

References

Acuna, J.A., and Y.C. Yortsos (1995). Application of fractal geometry to the study of networks of fractures and their pressure transient. *Water Resources Research*, **31(3)**, 527–540.

An, L.J. (1998). Development of fault discontinuities in shear experiments. *Tectonophysics*, **293**, 45–59.

Antonellini, M., and A. Aydin (1994). Effect of faulting on fluid flow in porous sandstones: petrophysical properties. *AAPG Bulletin*, **78(3)**, 355–377.

Antonellini, M., and A. Aydin (1995). Effect of faulting on fluid flow in porous sandstones: geometry and spatial distribution. *AAPG Bulletin*, **79(5)**, 642–671.

Aydin, A., and R.A. Schultz (1990). Effect of mechanical interaction on the development of strike-slip faults with echelon patterns. *Journal of Structural Geology*, **12**, 123–129.

Baecher, G.B., N.A. Lanney, and H.H. Einstein (1977). Statistical description of rock properties and sampling. *Proceedings of the 18th U.S. Symposium on Rock Mechanics*, A.I.M.E., 5C1:1–8.

Barton, C.A., and M.D. Zoback (1990). Self-similar distribution of macroscopic fractures at depth in crystalline rock in the Cajon Pass scientific drillhole. In: *Rock Joints*, Barton N., Stephansson O. (eds.), Balkema A.A., Rotterdam, Netherlands, 163–170.

Bour, O., and P. Davy (1999). Clustering and size distributions of fault patterns: Theory and measurements. *Geophysical Research Letters*, **26**, 2001–2004.

Bourgine, B., J.P. Chilès, and C. Castaing (1995). Simulation d'un réseau de fractures par un modèle probabiliste hiérarchique. *Cahiers de Géostatistique*, Fasc. 5, Ecole des Mines de Paris, 81–96.

Bourne, S.J., F. Brauckmann, L. Rijkels, B.J. Stephenson, A. Weber, and E.J.M. Willemse (2000). Predictive modelling of naturally fractured reservoirs using geomechanics and flow simulation. *Paper ADIPEC-0911*, Society of Petroleum Engineers, 10 p.

Bürgman, R., D.D. Pollard, and S.J. Martell (1994). Slip distribution on faults: effects of stress gradients, inelastic deformation, heterogeneous host-rock stiffness, and fault interaction. *Journal of Structural Geology*, **16**, 1675–1690.

Cacas, M.C., J. Letouzey, and W. Sassi (1997). Modélisation multi-échelle de la fracturation naturelle des roches sédimentaires stratifiées. *Comptes Rendus de l'Académie des Sciences de Paris*, t. 324, série II a, 663–668.

Caputo, R., and G. Santarato (1998). Extensional joints and faults: A 3D mechanical model for quantifying their ratio – Part 1: Theory. Part 2: Applications. In: *Mechanics of Jointed and Faulted Rock*, H.P. Rossmanith (ed.), A.A. Balkema, Rotterdam, Netherlands, 133–144.

Castaing, C., M.A. Halawani, F. Gervais, J.P. Chilès, A. Genter, B. Bourgine, G. Ouillon, J.M. Brosse, P. Martin, A. Genna, and D. Janjou (1996). Scaling relationships in intraplate fracture systems related to Red Sea rifting. *Tectonophysics*, **261**, 291–314.

Chilès, J.P. (1988). Fractal and geostatistical methods for modeling of a fracture network. *Mathematical Geology*, **20(6)**, 631–654.

Chilès, J.P. (1989). Three-dimensional geometric modelling of a fracture network. In *Geostatistical, Sensitivity, and Uncertainty Methods for Ground-Water Flow and Radionuclide Transport Modeling*, B.E. Buxton (ed.), Battelle Press, Columbus, Ohio, 361–385.

Chilès, J.P., and G. de Marsily (1993). Stochastic models of fracture systems and their use in flow and transport modeling. In *Flow and Contaminant Transport in Fractured Rock*, J. Bear, G. de Marsily, and C.F. Tsang (eds.), Academic Press, San Diego, California, Chap. 4, 169–236.

Christiansen, P.P., and D.D. Pollard (1997). Nucleation, growth and structural development of mylonitic shear zones in granitic rock. *Journal of Structural Geology*, **19(9)**, 1159–1172.

Conrad, F., and C. Jacquin (1973). Représentation d'un réseau bidimensionnel de fractures par un modèle probabiliste. Application au calcul des grandeurs géométriques des blocs matriciels. *Revue de l'I.F.P.*, **28(6)**, 843–890.

Crider, J.G., and D.D. Pollard (1998). Fault linkage: Three dimensional mechanical interaction between echelon normal faults. *Journal of Geophysical Research*, **103(24)**, 373–391.

Dershowitz, W.S. (1984). *Rock joint systems.* Ph. D. dissertation, MIT, Cambridge, Massachusetts.

Finkbeiner, T., C.A. Barton, and M.D. Zoback (1997). Relationships among in-situ stress, fractures and faults, and fluid flow: Monterey Formation, Santa Maria Basin, California. *AAPG Bulletin*, **81(12)**, 1975-1999.

Gauthier, B.D.M., M. Garcia, and J.M. Daniel (2002). Integrated fractured reservoir characterization: A case study in a North Africa field. *Paper SPE 79105*.

Genter, A., and C. Castaing (1997). An attempt to simulate fracture systems from well data in reservoirs. *International Journal of Rock Mechanics and Mining Sciences*, **34(3/4)**, p. 448, Paper No. 44. Full length paper on CD-ROM (K. Kim, ed.).

Gillespie, P.A., C.B. Howard, J.J. Walsh, and J. Watterson (1993). Measurement and characterisation of spatial distributions of fractures. *Tectonophysics*, **226**, 113–141.

Hatton, C.G., I.G. Main, and P.G. Meredith (1994). Non-universal scaling of fracture length and opening displacement. *Nature*, **367**, 160–162.

Jeyakumaran, M., J.W. Rudnicki, and L.M. Keer (1992). Modeling slip zones with triangular dislocation elements. *Seismic Society of America Bulletin*, **82**, 153–169.

Lantuéjoul, C., H. Beucher, J.P. Chilès, H. Wackernagel, and P. Elion (2004). Estimating the trace length distribution of fractures from line sampling data. This volume.

Lyman, G.J. (2003). Stereological and other methods applied to rock joint size estimation—Does Crofton's theorem apply? *Mathematical Geology*, **35(1)**, 9–23.

Martel, J.S., and W.A. Boger (1998). Geometry and mechanics of secondary fracturing around small three-dimensional faults in granitic rock. *Journal of Geophysical Research*, **103 B 9**, 21,299–21,314.

Odling, N.E. (1997). Scaling and connectivity of joint systems in sandstones from western Norway. *Journal of Structural Geology*, **19(10)**, 1257–1271.

Peacock, D.C.P. (1996). Field examples of variations in fault patterns at different scales. *Terra Nova*, **8**, 561–371.

Pohlmann, S., J. Mecke, and D. Stoyan (1981). Stereological formulas for stationary surface processes. *Mathematische Operationsforschung und Statistik*, **12(3)**, 429–440.

Priest, S.D., and J.A. Hudson (1976). Discontinuity spacings in rock. *International Journal of Rock Mechanics and Mining Sciences & Geomechanics Abstracts*, **13**, 135–148.

Renshaw, C., and D. Pollard (1999). Numerical simulation of fracture set formation: a fracture mechanics model consistent with experimental observations. *Journal of Geophysical Research*, **99**, 9359–9372.

Rives, T., and J.P. Petit (1990). Diaclases et plissements : une approche expérimentale. *Comptes Rendus de l'Académie des Sciences de Paris*, t. 310, série II, 1115–1121.

Schultz-Ela, D.D. (1988). Application of a three-dimensional finite-element method to strain field analyses. *Journal of Structural Geology*, **10(3)**, 263–272.

Schwandtke, A. (1988). Second order quantities for stationary weighted fibre processes. *Mathematische Nachrichten*, **139**, 321–334.

Segall, P., and D.D. Pollard (1983). Joint formation in granitic rock of the Sierra Nevada. *Geological Society of America Bulletin*, **94**, 563–575.

Sornette, A., P. Davy, and D. Sornette (1990). Growth of fractal fault patterns. *Physical Review Letters*, **65(18)**, 2266–2269.

Srivastava, R.M. (2002). Probabilistic discrete fracture network models for the Whiteshell research area. Ontario Power Generation, Report No: 06819-REP-01200-10071-R00, Toronto, Canada.

Stoyan, D., and H. Stoyan (1994). *Fractals, Random Shapes and Point Fields. Methods of Geometrical Statistics.* Wiley, Chichester, England.

Takayasu, H. (1985). A deterministic model of fracture. *Progress in Theoretical Physics*, **74**, 1343–1345.

Terzaghi, R.D. (1965). Sources of error in joint surveys. *Geotechnique*, **15(3)**, 287–303.

Thomas, A.L. (1993). *Poly3D: A three-dimensional, polygonal element, displacement discontinuity boundary element computer program with applications to fractures, faults, and cavities in the Earth's crust.* Masters dissertation, Stanford University, California.

Turcotte, D.L. (1986). A fractal model for crustal deformation. *Tectonophysics*, **132**, 261–269.

Turcotte, D.L. (1992). Fractal, chaos, self-organized criticality and tectonics. *Terra Nova*, **4**, 4–12.

Veneziano, D. (1978). *Probabilistic model of joints in rock.* MIT, Cambridge, Massachusetts.

Walmann, T. (1998). *Dynamics and Scaling Properties of Fractures in Clay-Like Materials.* Ph. D. thesis, University of Oslo, Norway.

Weinberger, R., V. Lyakhovsky, and A. Agnon (1999). Damage evolution and propagation paths of en-échelon cracks. In: *Rock Mechanics for Industry*, B. Amadei, Kranz, Scott, and Smeallie (eds.), Balkema A.A., Rotterdam, Netherlands, Vol. 2, 1125–1132.

Wen, R., and R. Sinding-Larsen (1997). Stochastic modeling and simulation of small faults by marked point processes and kriging. In *Geostatistics Wollongong '96*, E.Y. Baafi and N.A. Schofield (eds.), Kluwer, Dordrecht, Netherlands, Vol. 1, 398–414.

Willemse, E.J.M., D.C.P. Peacock, and A. Aydin (1997). Nucleation and growth of faults in limestones from Somerset, UK. *Journal of Structural Geology*, 1461–1477.

Yow, J.L. Jr. (1987). Blind zones in the acquisition of discontinuity orientation data. *International Journal of Rock Mechanics and Mining Sciences & Geomechanics Abstracts*, **24(5)**, 317–318.

Zhang, L., and H.H. Einstein (2000). Estimating the intensity of rock discontinuities. *International Journal of Rock Mechanics and Mining Sciences*, **37**, 819–837.

GEOSTATISTICAL SIMULATION OF FRACTURE NETWORKS

R. MOHAN SRIVASTAVA[1], PETER FRYKMAN[2] and MARK JENSEN[3]
[1] *FSS Canada Consultants, Toronto, Canada*
[2] *Danish Geological Survey, Copenhagen, Denmark*
[3] *Ontario Power Generation, Toronto, Canada*

Abstract. A geostatistical method has been developed by Ontario Power Generation to enable creation of 3D fracture network models (FNMs) that explicitly honor detailed information on surface lineaments. This approach provides a systematic and traceable method that is flexible and that accommodates data from many different sources. The detailed, complex and realistic models of 3D fracture geometry produced by this method serve as an ideal basis for developing rock property models to be used in flow and transport studies. These models are probabilistic in the sense that they consist of a family of equally likely renditions of fracture geometry. Such probabilistic models are well suited to studying issues that involve risk assessment and quantification of uncertainty.

The geostatistical procedure for simulating FNMs is described, and tested using field data collected from the Lägerdorf chalk quarry in northern Germany.

1 Introduction

Fractures play a dominant role in fluid flow and transport; 3D models of fracture networks can therefore be very useful as inputs to flow simulators. Existing procedures for simulating fracture network geometry typically simplify the undulating and curvilinear nature of fracture surfaces,approximating them as planar facets. Though this approximation is suitable for many flow and transport studies, it is unacceptable in situations where models of fracture geometry need to exactly honor known locations of a large, detailed and geometrically complex set of fractures.

Ontario Power Generation (OPG) is developing tools to assist the Deep Geologic Repository Technology Program (DGRTP) with integrating diverse types of data into geosphere models that are consistent with increasingly detailed surface and subsurface knowledge. One specific data integration task involves construction of 3D fracture network models that honor information available at the time of preliminary site characterization: surface lineaments, general structural geology principles, regional tectonic considerations and site-specific information on geomechanical characteristics. These various pieces of information provide constraints

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 295-304.

on fracture location: very strong constraints at surface in regions of good bedrock exposure; weaker constraints at depth and in regions of poor bedrock exposure.

A geostatistical simulation method has been developed for creating complex, detailed and realistic 3D fracture network models that honor the various pieces of information available for preliminary site characterization. It can also accommodate data that become available in later years, such as borehole data and information from other subsurface investigations. This method produces a family of equally probable renditions of 3D fracture geometry, each one different in detail, but all consistent with the same constraints.

2 Overview and implementation

Sequential gaussian simulation (SGS) is a procedure that is widely used for creating data-conditioned stochastic models of spatial phenomena. In a typical SGS study, the procedure is applied to volume-averaged rock properties (such as grades or permeabilities), and is performed at the nodes of a regular grid. In the procedure discussed here, SGS is applied to geometric attributes (strike of a fracture trace, or dip of a fracture surface). Furthermore, the locations at which SGS is applied are not nodes of a regular grid; instead, the procedure is applied at the tips of iteratively propagating fractures, which entails that the locations of the simulations nodes depend on the details of what has been simulated in previous iterations.

The original motivation for using an iterative procedure like SGS was to mimic the procedure proposed by Renshaw and Pollard (1994). Their approach to 2D fracture simulation was based on geomechanical principles, propagating fracture tips when stress at fracture tips exceeds a critical threshold. Their approach created very realistic synthetic images of fractures in a variety of stress environments, including many subtle features commonly observed in the field, such as zones of small *en echelon* features that bridge gaps between larger separate fractures.

Though undeniably successful, fracture simulation based on geomechanical principles proved to be prohibitively computationally intensive and has never been extended satisfactorily to 3D. By using the same broad approach — an iterative procedure for propagating fracture tips — but replacing geomechanical principles for fracture propagation with geostatistical rules, one is able to mimic much of the realism with less computational effort.

2.1 2D PROPAGATION OF SURFACE FEATURES

Figures 1 through 6 show the major steps in the first phase of the fracture simulation: 2D propagation of surface traces.

The procedure begins with the seeding of the initial fracture segments, each one of which is assigned an initial direction of propagation and an intended final length. For deterministic fractures that can be identified from aerial photography or surface reconaissance (the grey lines in Figure 1), the initial fracture segment is seeded at the midpoint, the initial direction of propagation is determined from the lineaments orientation at its midpoint and the intended final length is the length of the known fracture. In areas of poor bedrock exposure (the darker region on the left of Figure 1), additional fractures are seeded to bring the number of fractures up to the number predicted by the user-specified model of fracture density. These hidden fractures are assigned their locations according to a clustered Poisson process; their initial orientations are drawn from the model of azimuth distribution and their intended final lengths from the model of fracture length distribution. The filled dots in Figure 1 mark “constrained” endpoints, those whose propagation will be governed by the trace of known fractures. The open dots mark the “free” endpoints, those whose propagation will be determined by sequential gaussian simulation.

Figure 2 shows how free endpoints are propagated. These are visited in a random order and each one is propagated a fixed step length. The direction of propagation is determined by simple kriging using the closest segment of each nearby fracture as conditioning data. In the example shown in Figure 2, the azimuth data used in the kriging for the free endpoint marked by the question mark are shown in black. One additional piece of data (shown in grey) is included in the kriging: a point halfway to the nearest neighbor, with an orientation parallel to the line segment that connects the endpoint being propagated to its nearest neighbor. For the example in Figure 2, the distribution of possible azimuth directions has a mean of 70° and a standard deviation of 12°; this is shown by the pie-shaped slice centered on the mean and with an angular width of two standard deviations. A specific direction (the dashed line) is randomly drawn from the distribution and the endpoint is propagated in that direction.

With simple kriging being performed on the strike of the fracture traces, a variogram model is required for this strike attribute. For long features that are very nearly linear, the variogram of strike will show strong spatial continuity, typically modelled with quadratic behavior at short distances and with little or no nugget effect. For features that undulate, are kinked or meander, the variogram of strike will show less spatial continuity and may be modelled with a small nugget effect. The use of variogram information on the strike ensures realistic portrayal of the fractures’ undulations.

Figure 3 shows how constrained endpoints are handled differently. For simulated fractures that track deterministic features, their propagation simply follows the trace of the identified fracture. In the example in Figure 3, the constrained endpoint marked with the question mark is propagated along the trace of the deterministic trace shown in gray. As each endpoint is propagated, a new line segment is created. As in a conventional sequential simulation procedure, each newly simulated segment is available for use in all subsequent simple krigings.

When the propagation of a free endpoint collides with a previously simulated endpoint, that endpoint is terminated with a user-specified probability. Setting this truncation probability to 1 guarantees that when fractures meet, one of them will terminate, a pattern common with faults. Using lower probabilities allows fractures to cross each other, a pattern common with joints.

If both the free endpoints of a hidden fracture are terminated before the fracture reaches its intended length, then the remaining unused length is assigned to a fracture still being propagated. This preserves total fracture length in the study area, at the expense of increasing the variance of the fracture length distribution, a result of allowing more shorter fractures, as well as more longer ones.

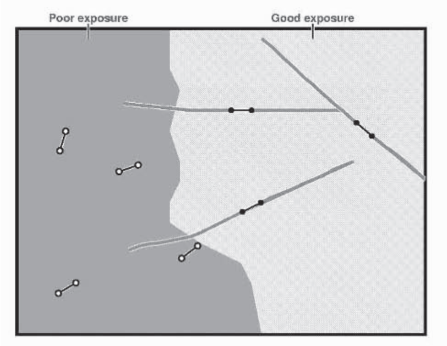


Figure 1. Initial seeding of the midpoints of the fracture traces.

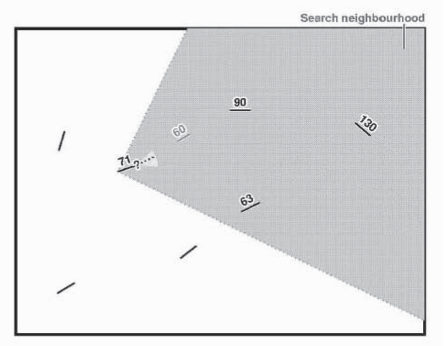


Figure 2. An example of the propagation of a free endpoint.

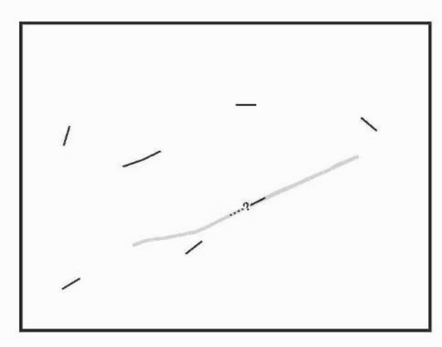


Figure 3. An example of the propagation of a constrained endpoint.

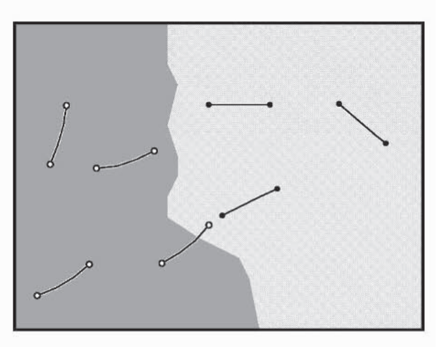


Figure 4. After one complete iteration through the endpoints.

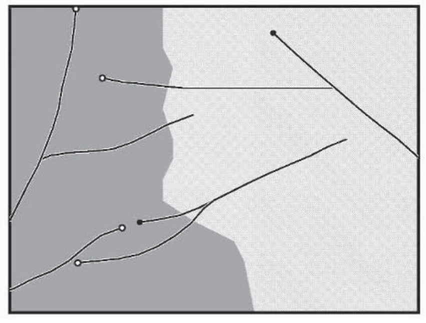


Figure 5. After five complete iterations through the endpoints.

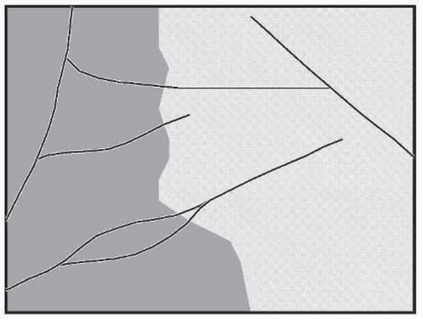


Figure 6. Final outcome of simulated fracture traces at surface.

Figure 4 shows the example after one complete iteration through all endpoints. As can be seen near the bottom of this figure, hidden fractures seeded in areas of poor exposure are allowed to propagate some user-specified distance into areas of good exposure.

Figure 5 shows the result after five complete iterations through all endpoints. Some of the fractures, such as the one whose initial propagation was shown in Figure 2, have reached their intended length and do not propagate any further. Others, such as the "hidden" one that first crossed into the area of good exposure in Figure 4, have terminated against other lineaments and do not propagate further. Some constrained endpoints, such as the eastern tip of the deterministic fracture noted in Figure 3, have been terminated because the lineament ends in an area of good bedrock exposure. Other constrained endpoints, such as the other end of the same feature, have passed into areas of poor bedrock exposure. When this happens, the endpoint becomes free and its propagation is governed by the SGS procedure since there is no longer a deterministic trace that can be followed.

Figure 6 shows the final result after several more iterations. The fractures originally identified in regions of good bedrock exposure have been honored; additional fractures have been added in regions of poor exposure; certain features have grown together into larger connected fracture systems; fractures truncate against each other in a plausible manner.

2.2 PROPAGATION TO DEPTH

Once the surface traces of the fractures have been simulated, the next step is the down-dip propagation of these traces to depth. This is accomplished using the same SGS procedure that was used to govern the 2D propagation of free endpoints. For the 2D propagation of surface traces, the geometric attribute being simulated was the strike direction of the fracture. For the down-dip propagation, the simulated attribute now becomes the dip of the fracture surface.

In the same way that each endpoint was propagated using SGS in 2D by simulating an appropriate strike direction, each line segment is now propagated down-dip, with SGS being used to simulate an appropriate dip. If the dip of a particular fracture is known (from surface reconnaisance measurements, for example), then this known dip can be used. The more common situation is that the dip is not known precisely, but only approximately. For example, regional tectonic considerations often imply that a certain set of fractures are likely the type of low-angle features commonly found in compressional environments; similarly, high-angle and sub-vertical features are usually more common in tensional environments.

When the dip of a particular fracture is not known precisely, SGS can still be guided to plausible values for the simulated dip by appropriate choices of the local mean used in the simple kriging, and by the sill of the variogram, which governs the kriging variance and the spread of the local conditional probability distributions of dip. Adjustments to the local mean can also be used to cause low-angle features to flatten with depth, a characteristic common of compressional "thrust" faults.

When there is no site-specific information on the fracture location at depth, as is typically the case for preliminary site characterization, all line segments are treated as “free” segments: they are not constrained to follow a specific down-dip trajectory. In applications where subsurface investigations do provide specific information on fracture location at depth, this type of data is easily accommodated in the same way that it was for the 2D propagation. The line segments corresponding to certain fractures can be treated as “constrained” and can be required to follow a specific trajectory.

2.3 PROPERTIES OF SIMULATED FRACTURE NETWORKS

Once the down-dip propagation of the fracture surfaces is complete, we have a complete 3D model of the geometry of the fracture network that:

- honors the surface traces of all identified fractures and, if available, subsurface information on the location of specific features;
- includes additional stochastic fractures in areas where poor exposure causes deterministic fractures to be under-represented;
- honors the fracture length distribution;
- honors the fracture orientation distribution; and,
- honors the user’s assumptions about how fractures truncate against one another.

Several applications of this method to different case study examples confirm that the resulting geometry is plausible both from a geomechanical point of view, as well as from a structural geology point of view (Srivastava, 2002; Sikorsky et al., 2002); one such recent case study is discussed below.

3 Lägerdorf case study

From December 1990 to May 1992, GEO-RECON, a Norwegian consulting company, undertook a fracture mapping program at the Lägerdorf quarry in northern Germany on behalf of the Joint Chalk Research Program, a multi-company research program that focused on subjects related to North Sea chalk reservoirs. Data collection procedures are described in detail by Koestler and Reksten (1992).

The data set consists of highly detailed maps of fracture traces for 12 parallel faces of one wall of the quarry as it was advanced in small increments (roughly 1 to 1.5 metres). Each face map spans a region approximately 230 metres long by 40 metres high. All sections are inclined at approximately 50°. The faces were available for mapping during production, where the excavation process continuously scrapes a layer of 1.5 m thickness off the quarry wall by an abrading conveyor belt. During the period when the 12 faces were mapped and described structurally, the quarry wall was advanced 25 metres. The set of 12 face maps therefore represents a high resolution $2\frac{1}{2}$D fracture data set of a 230×40×25m volume.

Figure 7 shows the face maps of the fractures on walls 7 through 11. There are three major directional sets:

1. A set with steep westerly dips; these are particularly dense near two major shear zones identified as S1 and S2 on the map for Wall 10 on Figure 7.

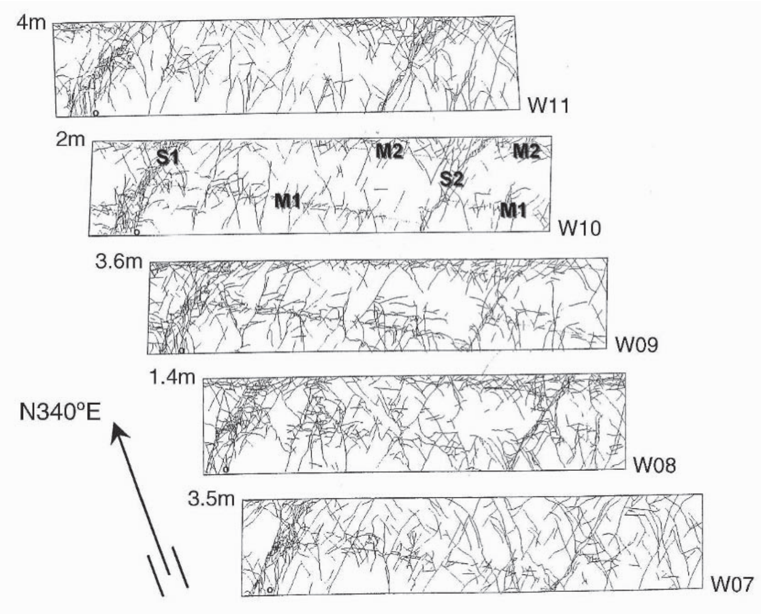


Figure 7. Fracture maps on walls 7 through 11.

2. A set with steep easterly dips; these are less common than the first set, but have similar lengths.
3. A set with shallow easterly dips; these are particularly dense along two marl layers identified as M1 and M2 on Figure 7. They tend to be shorter than the steeply dipping fractures. As shown on Figure 7, the major shear zones offset the marl layers.

To test the SGS-based geostatistical procedure for fracture simulation, the stack of parallel face maps was rotated so that the walls are essentially horizontal, with Wall 1 being at the top and Wall 12 being at the bottom. Following this rotation, we are able to treat the data from Wall 1 as a set of deterministic "surface" fractures that will be used as conditioning data for a simulation of 3D fracture geometry. In order to make the test as illuminating as possible, *all* data from Walls 2 through 12 were ignored. Once a set of simulations of 3D fracture geometry has been developed, each realization can be sliced along planes corresponding to the walls not used as conditioning data and the resulting simulated face map can be compared to the actual face map for each wall.

Figure 8 shows the conditioning data from Wall 1, the "surface" data for the test of the simulation procedure. In order to check the 2D propagation at surface,

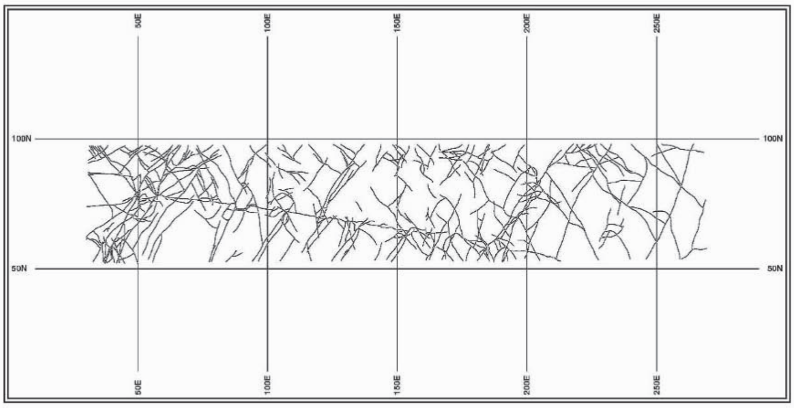


Figure 8. "Surface" fractures from Wall 1.

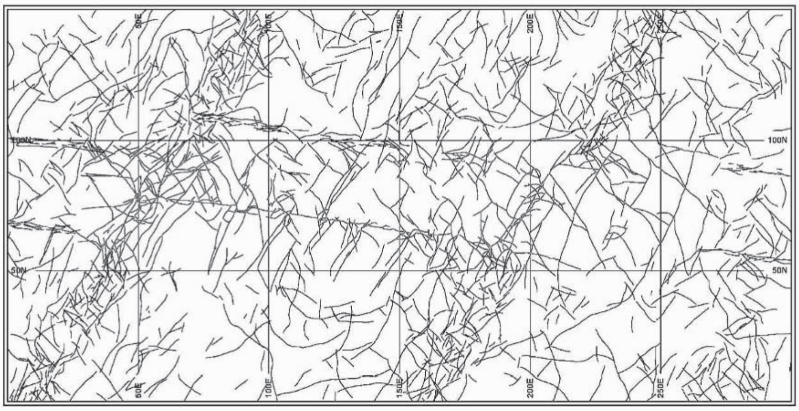


Figure 9. Realization No. 1 of simulated surface fractures.

the area covered by the mapped fractures on Wall 1 was extended to create a border, construed as a region of poor bedrock exposure where no deterministic fractures have been identified and where the simulation procedure will need to create a plausible rendition of hidden fractures.

Figure 9 shows one realization of the simulated traces at surface. The deterministic features identified by GEO-RECON are all honored, and pass quite seamlessly into the border region where the fracture traces are all simulated.

Figure 10 shows two realizations of simulated traces along the surface corresponding to Wall 11, along with the actual data from Wall 11. While the simulated fractures do have a plausible overall appearance, some of the details have been shifted slightly. The marl layers, for example, do not appear on the simulations at the same location as they do on the actual face map. The reasons for this discrepancy are well understood.

By strictly limiting the data available to the simulation to the measurements gathered from Wall 1, our model of the orientation of the marl layers (and of the

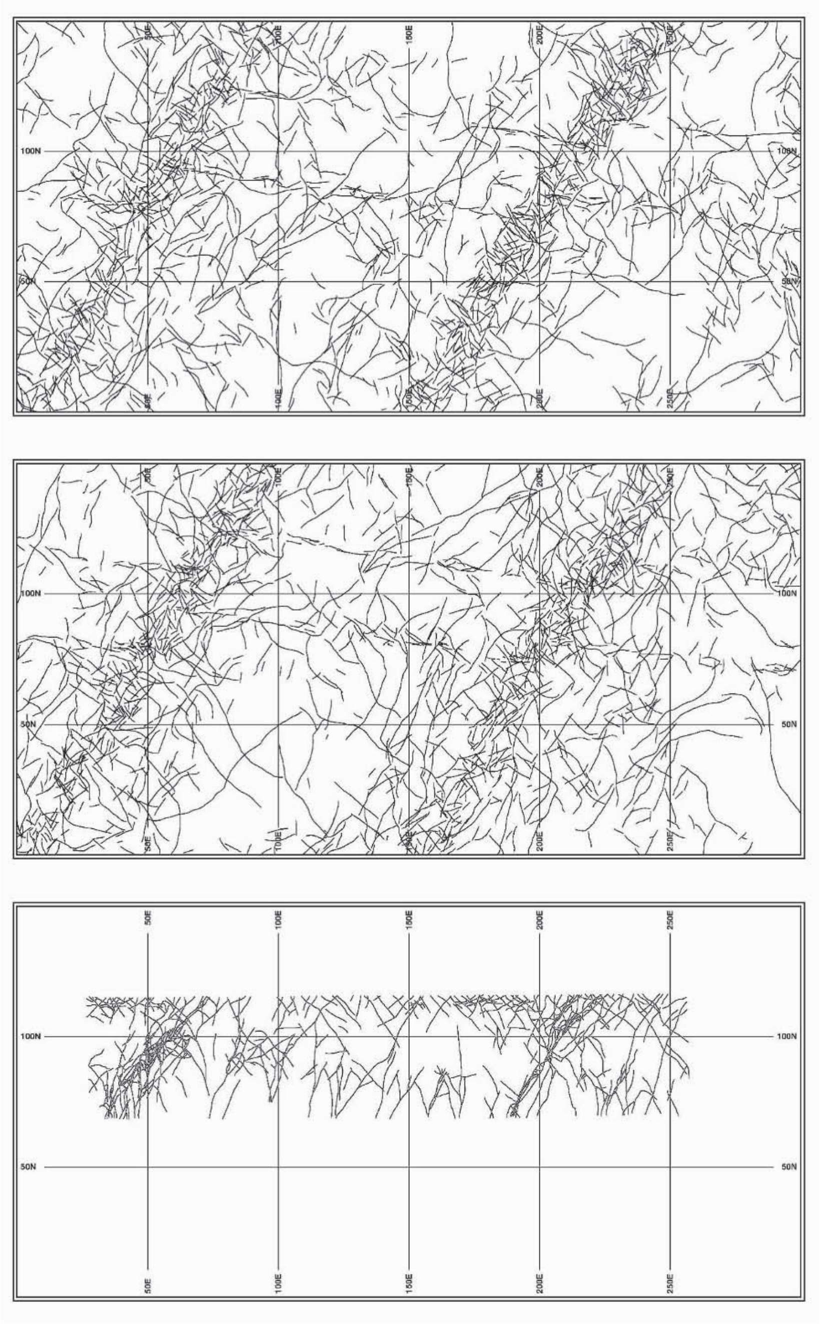


Figure 10. Simulated and actual fractures on Wall 11.

shear zones) is dependent entirely on the extrapolation of the planar orientation deduced from measurements available on Wall 1. The orientation of the marl layers deduced from Wall 1 data alone is slightly different than the orientation one can calculate from the full data set. So although the simulated fractures are definitely off in their depiction of the exact location of the densely fractured zones, this error would easily be corrected once subsurface investigations such as boreholes provided accurate 3D information on the location and orientation of the major geological features.

Future studies will compare the actual field data to the statistical and geometric characteristics of the clusters of simulated fractures. Flow simulation will also be used to compare actual and simulated flow-related characteristics, such as peak arrival times of injected tracer. For the moment, the preliminary results from the Lägerdorf case study are encouraging. The SGS-based procedure does create 3D fracture network models that are visually plausible and that honor complex and highly detailed surface information.

4 Conclusions

The proposed procedure for simulating fracture networks has now been tested on a variety of different problems at various scales. The Lägerdorf case study offers confirmation that the procedure does generate highly detailed and complex fracture patterns that mimic actual field observations.

Acknowledgements

The work presented in this paper has been funded by the Deep Geologic Repository Technology Program of Ontario Power Generation, and has benefitted greatly from the constructive criticism and encouragement of Jean-Paul Chilès, Alexandre Desbarats and Jaime Gomez-Hernàndez. The authors are also grateful to Andreas Koestler for his invaluable assistance with the data he collected and assiduously compiled from the Lägerdorf quarry.

References

Koestler, A.G., and Reksten, K., 1992, *3D geometry and flow behaviour of fractures in deformed chalk, Laegerdorf, Germany*, GEO-RECON Report no. 9012.126, GEO-RECON A.S, Oslo, Norway.

Renshaw, C., and Pollard, D., 1994, "Numerical simulation of fracture set formation: a fracture mechanics model consistent with experimental observations", *Journal of Geophysical Research*, vol. 99, pp. 9359–9372.

Sikorsky, R.I., Serzu, M., Tomsons, D., and Hawkins, J., 2002 *A GIS-based methodology for lineament interpretation and its application to a case study at AECL's Whiteshell Research Area in southeastern Manitoba*, Deep Geologic Repository Technology Program Report 06819-01200-10073, Ontario Power Generation, Toronto, Ontario.

Srivastava, R.M., 2002, *Probabilistic discrete fracture network models for the Whiteshell Research Area*, Deep Geologic Repository Technology Program Report 06819-01200-10071, Ontario Power Generation, Toronto, Ontario.

THE PROMISES AND PITFALLS OF DIRECT SIMULATION

OY LEUANGTHONG
Department of Civil & Environmental Engineering, 220 CEB,
University of Alberta, Canada, T6G 2G7

Abstract. The idea of direct simulation is to simulate in the space of the original data units, with minimal assumptions or transformations about the data distribution. A common approach to direct simulation is to proceed in a sequential fashion: direct sequential simulation (DSS). While the idea is not new, full development of the framework remains to be seen. The benefits of multiscale data integration, avoidance of the "Gaussian disease", and flexible distribution considerations are offset by problems with histogram reproduction, the pervasive influence of Gaussianity, and proportional effect reproduction.

This paper examines the promises and pitfalls of direct simulation with some illustrative examples, and also discusses the future of DSS as a practical alternative for natural resource characterization. The future of DSS calls for an engine other than kriging that accounts for possible dependency between the local variance and mean.

1 Introduction

Over the last decade, direct simulation has been proposed as a viable alternative to the venerable Gaussian simulation approaches. The idea of direct simulation is to simulate in the space of the original data units, with minimal assumptions or transformations about the data distribution. Behind this key idea is the principle of simple kriging.

Journel (1994) first showed that the covariance of simulated values reproduces the target covariance model *if* the simulated values are drawn from a distribution centred about the simple kriging (SK) mean and a variance given by the SK variance. Indeed, Bourgault (1997) showed this to be true for different distributional shapes including the uniform, dipole and of course, the Gaussian distribution. Caers (2000) also shows this for a uniform, double exponential, double exponential with a spike, and a "bootstrapped" distribution.

Covariance reproduction without relying on the Gaussian framework seeded the idea for direct simulation. The key premise for why direct simulation works lies in the orthogonality between the SK estimate, $Z^*(\mathbf{u})$, and the squared error which forms the basis for the SK error variance, $\sigma_{SK}^2(\mathbf{u})$. This can be thought of in terms of projections

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 305-314.

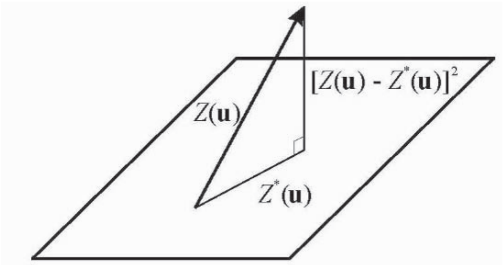


Figure 1 Kriging in terms of projection theory (redrawn from Journel and Huijbregts, 1978; Anton and Rorres, 1991).

where the squared error, $[Z(u)-Z^*(u)]^2$, is orthogonal to the space of all finite linear combinations of the random variables (RV), $Z(\mathbf{u}_\alpha)$, $\alpha = 1,..., n$ (Journel and Huijbregts, 1978) (see Figure 1). The kriging estimate, $Z^*(\mathbf{u})$, lies in this space as it is a linear combination of the RVs, $Z(\mathbf{u}_\alpha)$, $\alpha=1,..., n$:

$$Z^*(\mathbf{u}) = \sum_{\alpha=1}^{n} \lambda_\alpha Z(\mathbf{u}_\alpha) \tag{1}$$

The squared error term, $[Z(\mathbf{u})-Z^*(\mathbf{u})]^2$, represents the distance to the unknown true value, $Z(\mathbf{u})$. Based on Projection Theory, there is a unique and exact solution that yields the linear coefficients, λ_α, $\alpha=1, ..., n$, such that this distance is minimized (Journel and Huijbregts, 1978). This solution is referred to as the projection of $Z(\mathbf{u})$ onto this space. The corollary to kriging lies in the fact that the weights, λ_α, $\alpha=1, ..., n$, are determined such that the expected squared error, $E\{[Z(\mathbf{u})-Z^*(\mathbf{u})]^2\}$, is minimum. This visual interpretation of simple kriging can also be thought of as satisfying the Generalized Theorem of Pythagoras (Anton and Rorres, 1991).

Orthogonality of the kriged estimate and the squared error leads to an error variance that is independent of the data values, commonly referred to as the homoscedascity of kriging. Under a Gaussian paradigm, this poses no problems; in fact, it would be exactly right. In reality, natural phenomena rarely possess characteristics similar to the Gaussian distribution. This is particularly evident upon examining the relationship between the local average and the local variability, which, contrary to the homoscedasticity inherent in kriging, often reveals the presence of a strong relationship between the two statistics. This heteroscedastic relationship is more specifically known as the proportional effect (Journel and Huijbregts, 1978), and poses the most significant problem for direct simulation.

This paper presents the promises and pitfalls of direct simulation with some illustrative examples. Five main areas of discussion are highlighted: (1) principle of simple kriging, (2) implementation of direct simulation, (3) multiscale data integration, (4) histogram reproduction, and (5) accounting for the proportional effect. Finally, the future of DSS is discussed.

2 The Simple Kriging Principal

Reproduction of the covariance only requires that the conditional probability distributions have a mean and variance given by simple kriging (Journel, 1994). Journel proved this by showing firstly, the detailed simulation of a variable at location **u**, then adding this simulated value to simulate the next location, **u**', and finally checking the covariance between these two simulated variables.

Consider a stationary random variable, Z(**u**), with zero mean and unit variance. Firstly, construct a simulated value such that it can be decomposed as

$$Z_s(\mathbf{u}) = m(\mathbf{u}) + R_s(\mathbf{u})$$

where $m(\mathbf{u})$ is the expected value at location $\mathbf{u} \in$ domain, A, and $R(\mathbf{u})$ is a random variable drawn from a distribution with zero mean and variance, $\sigma^2(\mathbf{u})$. The local mean is given by the kriging mean (Equation 1), and the variance is given by the SK variance:

$$\sigma_{SK}^2(\mathbf{u}) = 1 - \sum_{\alpha=1}^{N} \lambda_\alpha C(\mathbf{u} - \mathbf{u}_\alpha) \tag{2}$$

where C($\mathbf{u}$ - $\mathbf{u}_\alpha$) is the covariance between the location **u** and the data located at $\mathbf{u}_\alpha$, α=*1,..., n*, $\sigma_{SK}^2(\mathbf{u})$ is the simple kriging variance, and the weights, λ_α, α=*1,...,n* are obtained by solving the normal equations:

$$\sum_{\beta=1}^{n} \lambda_\alpha C(\mathbf{u}_\beta - \mathbf{u}_\alpha) = C(\mathbf{u} - \mathbf{u}_\alpha), \quad \alpha = 1,...,n \tag{3}$$

This simulated value is added to the conditioning data set, and simulation is performed at the next location $\mathbf{u}'$=$\mathbf{u}_{n+1}$ with the following kriged mean and variance:

$$Z^*(\mathbf{u}') = \sum_{\alpha=1}^{n} \lambda_\alpha z(\mathbf{u}_\alpha) + \lambda_{n+1} Z_s(\mathbf{u}) \tag{4}$$

$$\sigma_{SK}^2(\mathbf{u}') = 1 - \sum_{\alpha=1}^{n} \lambda_\alpha C(\mathbf{u}' - \mathbf{u}_\alpha) - \lambda_{n+1} C(\mathbf{u}' - \mathbf{u}) \tag{5}$$

Note that the weights λ_α, α=*1,...,n+1* are *not* the same as the weights λ_α, α=*1,...,n* obtained from solving the system in Equation 3. The simulated value is given as

$$Z_s(\mathbf{u}') = Z^*(\mathbf{u}') + R_s(\mathbf{u}')$$

The covariance between the two simulated variables is then examined:

$$\begin{aligned} C(\mathbf{u} - \mathbf{u}') &= E\{Z_S(\mathbf{u}) \cdot Z_S(\mathbf{u}')\} \\ &= E\{Z^*(\mathbf{u}) \cdot Z^*(\mathbf{u}')\} + E\{Z^*(\mathbf{u}) \cdot R_S(\mathbf{u}')\} + \\ &\qquad E\{Z^*(\mathbf{u}') \cdot R_S(\mathbf{u})\} + E\{R_S(\mathbf{u}) \cdot R_S(\mathbf{u}')\} \end{aligned} \tag{6}$$

where $E\{Z^*(\mathbf{u})\cdot R_S(\mathbf{u}')\}$ and $E\{R_S(\mathbf{u})\cdot R_S(\mathbf{u}')\}$ are zero since $Z^*(\mathbf{u})$ and $R_S(\mathbf{u}')$ are independent of each other and $R_S(\mathbf{u})$ and $R_S(\mathbf{u}')$ are also independent. The remaining portions of the right hand side are non zero since the kriged mean at the second location depends on the mean and randomly drawn value at the first location.

Expanding and simplifying the remaining two terms yields

$$E\{Z^*(\mathbf{u}')\cdot Z^*(\mathbf{u})\} = \sum_{\beta=1}^{n} \lambda_\beta C(\mathbf{u}_\beta - \mathbf{u}) + \lambda_{n+1}\left[1-\sigma_{SK}^2(\mathbf{u})\right] \tag{7}$$

$$E\{Z^*(\mathbf{u}')\cdot R_S(\mathbf{u})\} = \lambda_{n+1}\sigma_{SK}^2(\mathbf{u}) \tag{8}$$

Equations 7 and 8 are substituted into Equation 6:

$$\begin{aligned} C(\mathbf{u}-\mathbf{u}') &= \sum_{\beta=1}^{n} \lambda_\beta C(\mathbf{u}_\beta - \mathbf{u}) + \lambda_{n+1}\left[1-\sigma_{SK}^2(\mathbf{u})\right] + \lambda_{n+1}\sigma_{SK}^2(\mathbf{u}) \\ &= \sum_{\beta=1}^{n} \lambda_\beta C(\mathbf{u}_\beta - \mathbf{u}) + \lambda_{n+1} \\ &= C(\mathbf{u}'-\mathbf{u}) \end{aligned}$$

It is by this logic that Journel (1994) proved that so long as the conditional mean and variance are provided by simple kriging, covariance reproduction could be achieved without making any assumptions about the distributional shape. This is an exciting result as it opened the way for geostatisticians to consider simulation outside of the Gaussian framework and without the inference effort required under the indicator paradigm.

3 DSS Methodology

A common approach to simulation is to proceed in a sequential fashion; thus, Direct Sequential Simulation (DSS) was coined (Xu and Journel, 1994). The sequential simulation framework is straightforward:

1. Select a random path visiting all locations.
2. At each location:
   a. Search for all nearby data of different types and/or scale and previously simulated nodes (e.g. *P* data types with n_p samples).
   b. Perform simple kriging to determine the parameters corresponding to the conditional distribution, $F\left(Z(\mathbf{u})\middle| Z_p(\mathbf{u}_1),\ldots,Z_p(\mathbf{u}_{n_p})\right)$, *p=1, ..., P.*
   c. Draw a simulated value from this conditional distribution using Monte Carlo simulation. This simulated value is added to the conditioning data set.
3. Proceed to next node and repeat Step 2, until all locations are simulated.

The virtues of simplicity cannot be understated. The sequential algorithm was proposed by Johnson (1987), and is common in most geostatistical literature (Isaaks, 1990; Goovaerts, 1997; Deutsch and Journel, 1998; Chilès and Delfiner, 1999; Sinclair and Blackwell, 2000). There are other approaches for simulation, including the matrix approach (Davis, 1987) and turning bands (Journel and Huijbregts, 1978); however, the simplicity and efficiency of sequential simulation has made it the most popular approach in practice.

Indicator and Gaussian simulation have long been the "standard" geostatistical methods of choice in modern practice. Unlike sequential Gaussian simulation and sequential indicator simulation, the promise of DSS is that neither pre- nor post-processing steps are required. There is no need for data transformation, whether it is to a Gaussian or an indicator formalism. This sequential approach is common in mainstream numerical modelling, regardless of whether that modelling is performed under a parametric or non-parametric model.

4 Multi-Scale Data Integration

The current motivation for development of the direct simulation framework is the promise of integrating multiple data types from different sources and of different scales. Integrating data of different volume supports is neither new nor difficult in theory. Cokriging using average covariance/variograms permits consideration of multiscale data that average linearly. In fact, the generalized cokriging equations are straightforward to obtain.

Consider P stationary random variables, Z_p, $p=1,...,P$ with mean μ_p defined on support V_p centred at location $\mathbf{u}_{\alpha p}$, where $\alpha = 1,..., n_p$ and n_p is the number of available data of type p. It is not necessary that the volume supports V_p, $p=1,...,P$ be constant.

$$Z(\mathbf{u}_{\alpha p}) = \frac{1}{V_p} \int_{V_p} Z_p(\mathbf{u}_{\alpha p}) du$$

Without loss of generality, consider the residual of Z_p, $Y_p = Z_p - \mu_p$. Simple cokriging of the residual yields the following simple cokriging (SCK) variance:

$$\sigma^2_{SCK} = \overline{C}\big(V_i(\mathbf{u}), V_i(\mathbf{u})\big) - \sum_{p=1}^{P} \sum_{\alpha=1}^{n_p} \lambda_{\alpha p} \overline{C}\big(V_i(\mathbf{u}), V_p(\mathbf{u}_{\alpha p})\big)$$

where

$$C(V_i, V_j) = \frac{1}{|V_i||V_j|} \int_{V_i} dy \int_{V_j} C(y - y') dy'$$

and the weights are determined by simultaneously solving the $\sum_{p=1}^{P} n_p$ equations that constitute the simple co-kriging system of equations:

$$\sum_{p'=1}^{P}\sum_{\beta=1}^{n_{p'}}\lambda_{\beta p'}\overline{C}\Big(V_p(\mathbf{u}_{\alpha p}),V_{p'}(\mathbf{u}_{\beta p'})\Big)=\overline{C}\Big(V_i(\mathbf{u}),V_p(\mathbf{u}_{\alpha p})\Big),\quad p=1,\ldots,P \tag{9}$$

The resulting cokriging estimate and estimation variance correspond to the conditional expectation and variance of the RV $Y_p(\mathbf{u})$.

Greater efficiency can be achieved by simultaneously cokriging M multiple data types, where $M \leq P$. This is simply achieved by changing the column vector of weights and right hand side covariance into an M x P matrix. An additional index is required to indicate the variable to be estimated. For this purpose, the *m*, *m=1, ..., M*, index is introduced.

$$Y_i^*(\mathbf{u})=\sum_{p=1}^{P}\sum_{\alpha=1}^{n_p}\lambda_{\alpha p}^{1}Y_p(\mathbf{u}_{\alpha p})$$

$$\vdots$$

$$Y_M^*(\mathbf{u})=\sum_{p=1}^{P}\sum_{\alpha=1}^{n_p}\lambda_{\alpha p}^{M}Y_p(\mathbf{u}_{\alpha p})$$

Solving for the weights of the resulting co-kriging system requires little additional effort since the large left hand side data to data covariance matrix (in Equation 9) only has to be inverted once. Matrix multiplication of the inverted covariance matrix with the additional *M-1* columns of the right hand side covariance will give the weights to estimate the other *M*-1 additional variables. In fact, most solvers can be modified to solve systems of simultaneous equations with multiple right hand sides without explicitly solving for an inverse. The only additional computation required in order to simultaneously estimate the collocated data types is the determination of the right hand side volume to volume covariance between the location to be estimated and the nearby data of *P* types.

While cokriging of one variable gives the conditional expectation and variance of the RV, simultaneous cokriging of multiple RVs gives the conditional mean vector and covariance matrix of the *M* RVs. Simulation using these distributional parameters must still be performed.

All this is fine in the context of estimation where cokriging can be performed in the space of the data; however, in the context of Gaussian simulation, which is the most common practical simulation method, using average statistics after a non-linear transformation and back transforming to original units, does not work. Consider three numbers: 1, 2 and 10. The average of these three numbers is 4.33. Now consider an exponential transform, e^x where x is the data. This transform gives: 2.718, 7.389 and 22026.470, respectively. The average of the transformed values is 7345.524, which after back transformation yields 8.902. This is clearly not the same as the average in original space. Thus averaging in a non-linear space, such as Gaussian space, does not provide an appropriate method of accounting for multiscale data. This provides, yet, another impetus for pursuing DSS.

5 Histogram Reproduction

The topic of histogram reproduction is quite broad. It not only encompasses the obvious global distribution reproduction, but it also addresses the challenge of inferring the local distribution based on only two parameters. While this is sufficient information for a parametric model like the Gaussian model, it is often inadequate for more realistic non-parametric distributions.

The lack of a distributional assumption requirement is an obvious benefit for DSS. Natural phenomena rarely follow a parametric form such as the Gaussian distribution, and while quantile transformation permits a change from one distribution to any another, there is nothing that says we *should* transform the data to a parametric form. That data transformation is a widely accepted part of the modelling work flow speaks volumes about our strong and continued reliance on simple, yet restrictive mathematical models.

In fact, one could argue that the effect of data transformation on the true spatial distribution of the data may be undesired. Transformation to and back-transformation from Gaussian space yields some disturbing results when applied to skewed distributions. While statistical fluctuations are an inherent property of stochastic simulation, it is expected that these deviations should be reasonable and unbiased. For any one realization, minor fluctuations from a zero mean and unit variance are expected; however, when these values are back transformed to original units a slight shift in the mean in normal space may translate to a more significant shift of the mean in original units. Similarly, the combined fluctuation of the mean and variance in normal space may translate to more noticeable shifts in original space. This is particularly true for skewed distributions, which is the case for some real phenomena. Fixes to this particular problem have been proposed (Journel and Xu, 1994); yet this can be avoided altogether if we do not perform any data transformation prior to modelling – hence direct simulation.

Although Journel (1994) showed that covariance reproduction was achievable without any distributional assumptions, histogram reproduction remained a challenge. Most of the last decade has seen the majority of research focussed on this specific issue in DSS. Soares (2001) proposed to determine the local cumulative distribution function (cdf) by sampling from part of the global cdf. Caers (2000) suggested the use of a posterior correction of the histogram originally proposed by Journel and Xu (1994), in combination with an acceptance/rejection approach to determining the local cdf. Oz et. al. (2003) proposed the prior use of a Gaussian transform to determine a table of local distributions that could be accessed during DSS.

Despite the fact that DSS permits different shapes of the local distributions, the global distribution of simulated values tend to a symmetric, bell-shaped distribution characteristic of the Gaussian distribution (see Bourgault (1997) and Caers (2000)). This is a reflection of the pervasive influence of the Central Limit Theorem, sometimes referred to as the "Gaussian disease". Of the different approaches to infer the local distribution, only the approaches proposed by Soares (2001) and Oz et.al. (2003) are successful at reproducing the global distribution without need for a post-simulation histogram correction.

While histogram reproduction is key to the success of any simulation approach, this is not a significant obstacle in the widespread consumption of DSS. Actually, the work conducted in the past decade shows that there are any number of tricks and tools that can be employed to reproduce the histogram with varying degrees of desire. Although we intuitively understand that different distributions should exist to reflect different local regions, there is nothing in the prevailing DSS algorithms that will account for the proportional effect. The practicality and hence, viability, of DSS depends heavily on the promise of honouring the proportional effect.

6 Proportional Effect

By virtue of DSS' dependence on kriging, the resulting local variance is independent of the data values and the estimate, hence it is homoscedastic. In contrast, the variance of mineral grades or petrophysical properties found in a real deposit or reservoir often changes depending on the local mean – a property called heteroscedasticity. For example, it is common to find a low variance in a low valued area, and a correspondingly high variance in a high valued area. This heteroscedastic behavior is commonly referred to as the proportional effect (Journel and Huijbregts, 1978).

Consider the well-known Walker Lake data set and the lead pollution data from Dallas. A moving average approach was used with non-overlapping windows to determine the relationship between the local mean and variance. Figure 2 shows a very strong positive correlation for both data sets, in fact its relation appears quadratic, i.e.

$$\sigma^2(\mathbf{u}) = f\left(m(\mathbf{u})^2\right)$$

Note that this relationship is characteristic of real data (alternatively, it is sometimes shown as a linear relation between the standard deviation and the mean value), and it is more pronounced for a lognormal distribution (Armstrong (1998), Chilès and Delfiner (1999)). This relationship is neither new nor surprising. Journel and Huijbregts (1978), Isaaks and Srivastava (1991), Goovaerts (1997), and Chilès and Delfiner (1999), have all discussed the importance of the proportional effect in natural resource characterization. It is precisely in this aspect that direct simulation presents its biggest promise.

Yet there is a major flaw in the foundation of direct simulation. Its basis is founded in kriging, which yields a local variance that is data-value independent. As a result, it cannot produce models that will reproduce the heteroscedastic behaviour that would otherwise be found in real mineral deposits or reservoirs. Clearly, the flaw lies in the very fact that least squares estimation is the engine behind the simulation. For it to fulfil its promise, direct simulation must be built on a method that yields dependent mean and variance.

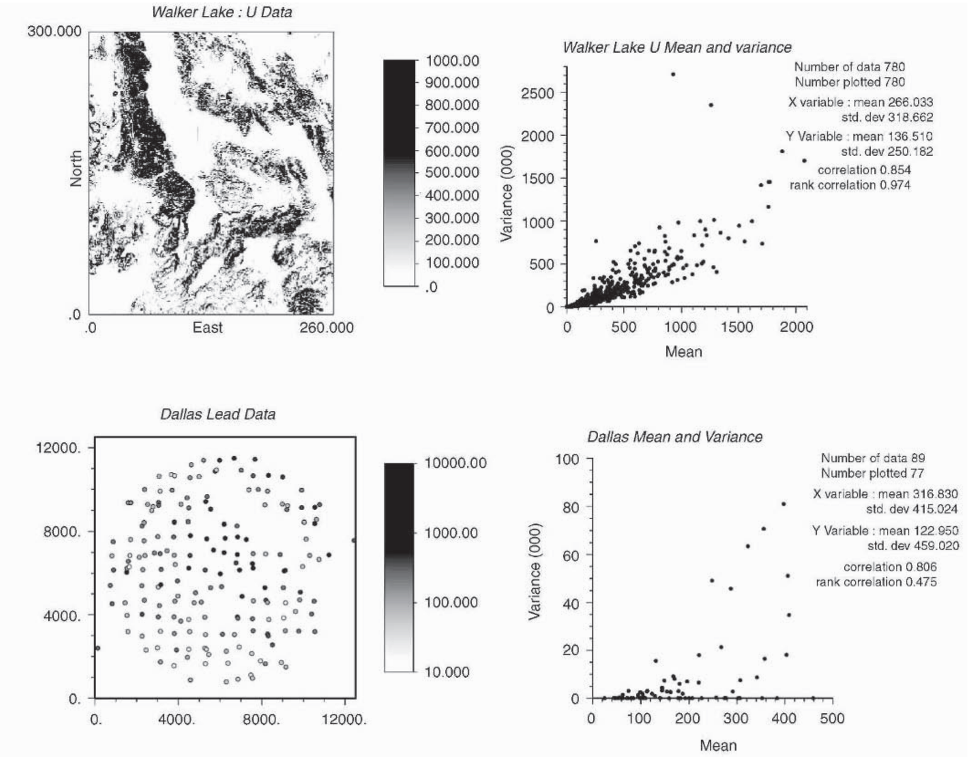


Figure 2 Illustration of proportional effect for Walker Lake data (top), and the lead pollution data from Dall (bottom). Plan view of the data is shown on the left, and crossplots of local variance vs. local mean are shown on the right.

7 Future of DSS

DSS presents one of the future avenues for geostatistics. It is among the latest in a series of simulation approaches that have been introduced in the last two decades. Whether it will rank among the “standard” approaches remains to be seen, advances in particular areas will certainly be key to its popularity. DSS promises (1) the ability to integrate multiple scale data since no transformation of the data is required, (2) reduced reliance on the multiGaussian paradigm, (3) simplicity in methodology, and (4) flexibility to consider different local distribution shapes to account for multivariate non-stationarity.

These promises, however, are balanced by the pitfalls of DSS which include (1) the unavoidable influence of multiGaussianity due to the Central Limit Theorem, (2) problems in histogram reproduction which have led to ad hoc post-processing techniques, (3) the inability to account for spatial heteroscedasticity, specifically the proportional effect, and (4) flexibility in using different distribution shapes locally has not been shown to be practically advantageous or straightforward to implement.

A number of issues must still be resolved to show a real advantage to DSS. The practical significance of accounting for the proportional effect is enormous. Resolution of this issue will lend serious credibility to DSS in construction of realistic numerical models, for application in all natural resource sectors. A second area of research lies in inference of the multivariate distribution. Many authors have expended tremendous research energies into univariate distribution inference, yet the true multiscale data integration benefits of direct simulation will never be realized if the multivariate distribution cannot be properly inferred.

Although DSS was built on the principles of simple kriging, its future cannot remain anchored to simple kriging. It does not lie in the homoscedastic kriging variance, as real data show a very strong relationship exists between the variance and the data values. For it to be of practical significance and in fact, to prevent it from simply becoming an academic exercise, the underlying principle of DSS must permit a heteroscedastic variance that is data-value *dependent*. This is contrary to its simple kriging foundations.

References

Armstrong, M., *Basic Linear Geostatistics*, Springer-Verlag, 1998.

Bourgault, G., Using Non-Gaussian Distributions in Geostatistical Simulations, *Mathematical Geology*, vol. 29, no. 3, 1997, p. 315-334.

Caers, J., Adding Local Accuracy to Direct Sequential Simulation, *Mathematical Geology*, vol. 32, no. 1, 2000, p. 815-850.

Chiles, J.P. and Delfiner, P., *Geostatistics: Modeling Spatial Uncertainty*, John Wiley & Sons, 1999.

Davis, M.W., Production of Conditional Simulations via the LU Triangular Decomposition of the Covariance Matrix, *Mathematical Geology*, vol. 19, no. 2, 1987, p. 91-98.

Deutsch, C.V., *Geostatistical Reservoir Modeling*, Oxford University Press, 2002.

Deutsch, C.V. and Journel, A.G., *GSLIB: Geostatistical Software Library and User's Guide*, 2nd Edition, Oxford University Press, 1998.

Goovaerts, P., *Geostatistics for Natural Resources Evaluation*, Oxford University Press, 1997.

Isaaks, E.H., *The Application of Monte Carlo Methods to the Analysis of Spatially Correlated Data*, PhD thesis, Stanford University, Stanford, CA, USA, 1990.

Johnson, M.E., *Multivariate Statistical Simulation*, John Wiley & Sons, 1987.

Journel, A.G., Modeling Uncertainty: Some Conceptual Thoughts, *Geostatistics for the Next Century*, R. Dimitrakopoulos, ed., Kluwer Academic Publishers, 1994a.

Journel, A.G. and Xu, W., Posterior Identification of Histograms Conditional to Local Data, *Mathematical Geology*, vol. 22, no. 3, 1994b, p. 323-359.

Journel, A.G. and Huijbregts, C.J., *Mining Geostatistics*, Academic Press, 1978.

Oz, B., Deutsch, C.V., Tran, T.T. and Xie, Y., DSSIM-HR: A Fortran 90 Program for Direct Sequential Simulation with Histogram Reproduction, *Computers & Geosciences*, vol. 29, 2003, p. 39-51.

Sinclair, A.J. and Blackwell, G.H., *Applied Mineral Inventory Estimation*, Cambridge University Press, 2002.

Soares, A., Direct Sequential Simulation and Cosimulation, *Mathematical Geology*, vol. 33, no. 8, 2001, p. 911-926.

SAMPLE OPTIMIZATION AND CONFIDENCE ASSESSMENT OF MARINE DIAMOND DEPOSITS USING COX SIMULATIONS

GAVIN BROWN
De Beers Group Services, Boundary Terraces, Mariendahl Lane, Newlands, 7725, Cape Town, South Africa

CHRISTIAN LANTUÉJOUL
Ecole des Mines, Centre de Géostatistique, 35 rue Saint-Honoré, 77305, Fontainebleau, France

CHRISTIAN PRINS
De Beers MRM R&D, Mendip Court, Bath Road, Wells, BA53DG, United Kingdom

Abstract. Conditional simulations are used to develop realistic and quantitative images of spatial variability for analysis. In particular, they are often used to evaluate the impact of uncertainty in applying economic optimization to a natural resource. Algorithms for conditional simulation of geological data and ore grades are well established. However, the majority are not applicable to placer diamond deposits which exhibit a discrete and patchy textural nature. This prompted the development of a new, enhanced suite of algorithms specifically designed for conditional simulations of diamond deposits. The Cox simulation algorithm is fast, incorporates complex mineralization structures, handles large sets of conditioning data and covers several different discrete distributions including those that are highly skewed. In addition, tests are available for model validation of both the distributional and spatial integrity. This enhanced conditional simulation tool is used, firstly, to determine the confidence limits of block estimates and secondly, to quantify and manage the risk of selective mining above an economic cut-off grade.

1 Introduction

Marine diamond evaluation is a very challenging process, balancing the sampling strategy, the estimation technique and the exploitation scenario. These challenges can be investigated by simulating a flexible model that can accommodate many specific features often observed in diamond samples, such as a long distributional tail and high occurrence of zero values.

This model is the so-called Cox process. Since its inception for modeling diamond deposits (Kleingeld, 1987), considerably more data has been collected in different

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 315-324.

marine environments, describing a large variety of distributional features for simulations to be generated. This prompted the requirement for enhancement of the Cox process. With the ability to perform conditional Cox simulations, the efficiency of sampling campaigns in different geological environments can be investigated. Furthermore, their impact on the assessment of spatial risk can be tested. This will be illustrated in two documented examples.

2 The Cox model

2.1 DEFINITION

Assume that the deposit is partitioned into a family of small congruent domains $(v_i,\, i \in I)$. For each index $i \in I$, let N_i denote the random number of particles falling within the domain v_i. In this paper, a model for the deposit is specified by the multivariate distribution of the random vector $(N_i,\ i \in I)$. The specification implies that any mineralization structure that is present at a smaller scale than the domains, is not accounted for by the model. This is compatible with the data collected, as the sample information does not comprise of the coordinates of its particles, but only the number of particles within each sample. In practical applications, the size and shape of the domains equate to those of the samples.

The number of particles in a domain v_i is affected by several factors, including the particle source, the local terrane and footwall lithology. As it is generally impossible to separately assess the contribution of each factor, it is convenient to summarize them as a single factor, say Z_i, that represents the propensity of the domain v_i to be rich. The larger the factor Z_i, the greater the chance that v_i will contain many particles. For this reason, Z_i is named the potential of the domain v_i.

For the Cox model, the number of particles within each domain is independently Poisson distributed, with the mean of each domain equal to its potential. However, as the potentials of the domains are unknown, they are considered random. Thus the numbers of particles within the domains are not independent - neighboring domains have correlated potentials - but are only conditionally independent. In this paper, we let $Z_i = \varphi(Y_i)$ for each $i \in I$, where $(Y_i, i \in I)$ is a standardized Gaussian vector.

2.2 STATISTICAL INFERENCE

The spatial distribution of the Cox model is characterized by two parameters, namely the anamorphosis φ of the potential of the domains and the covariance C of the underlying Gaussian vector Y. The inference of each parameter is considered in turn.

By definition, φ is determined by the distribution of Z_i and as there is a one-to-one correspondence between the distributions of Z_i and N_i (Feller, 1971), φ is also determined by the distribution of N_i. The statistical inference of φ can therefore be reduced to that of N_i.

In practice, the distribution of N_i is often chosen within a family of pre-specified models (negative binomial, Sichel, Cox-lognormal...), and its statistical inference

consists of estimating the parameters of the selected model from experimental statistics provided by the data. Such parameters include the mean number of particles per sample, the variance and the proportion of barren samples. For example two models are presented with their corresponding potential models:
– if N_i follows a negative binomial distribution with index $\alpha > 0$ and parameter $p = 1 - q$, then Z_i is gamma distributed with the same index α and scale factor $b = q/p$:

$$p_n = \frac{\Gamma(\alpha + n)}{\Gamma(\alpha)\, n!} q^\alpha p^n \quad n \in \mathbb{N} \qquad\qquad f(z) = \frac{b^\alpha}{\Gamma(\alpha)} e^{-bz} z^{\alpha-1} \quad z \in \mathbb{R}^+$$

– if N_i follows a Sichel distribution[1] with index $\alpha > 0$ and tail parameter $0 < \theta < 1$ (Sichel, 1973), then Z_i follows an inverse Gaussian distribution with parameters $a = (1-\theta)/\theta$ and $b = \alpha^2\theta/4$:

$$p_n \propto \frac{(\alpha\theta/2)^n}{n!} K_{n-1/2}(\alpha) \quad n \in \mathbb{N} \qquad\qquad f(z) \propto \frac{1}{z^{3/2}} \exp\left(-az - \frac{b}{z}\right) \quad z \in \mathbb{R}^+$$

Once φ has been estimated, several procedures can be performed to estimate the covariance C of Y. The simplest one rests on the fact that the covariances of N and Z differ only by their nugget effect:

$$Cov\{N_i, N_j\} = Cov\{Z_i, Z_j\} + E\{N_i\} 1_{i=j}$$

The covariance of Z can easily be derived from that of Y when φ has been expanded into Hermite polynomials:

$$Cov\{Z_i, Z_j\} = \sum_{n=1}^{\infty} \frac{\varphi_n^2}{n!} C_{ij}^n$$

Combining both equations and assuming that $Cov\{N_i, N_j\}$ is experimentally accessible, we can obtain C_{ij} from the equation

$$Cov\{N_i, N_j\} = \sum_{n=1}^{\infty} \frac{\varphi_n^2}{n!} C_{ij}^n + E\{N_i\} 1_{i=j}$$

This equation returns a single solution, due to the fact that its right-hand side member is a monotonic increasing function.
When all values for C_{ij} have been estimated, the subsequent modeling of C is necessary to ensure it is positive definite.

2.3 SIMULATION

The simulation using the Cox process with anamorphosis φ, covariance C and incorporating the available data comprising of a family of samples or blocks,

[1] In this formula, K_μ stands for the modified Bessel function of order μ.

is presented. Each sample and block is represented by a subset of indices of I (consisting of only one index in the case of a sample). The conditioning data can be written as $\big(N_A = n_A, A \in \mathcal{A}\big)$ for any family $\mathcal{A}$ of subsets of I. There is no inconvenience in assuming that the supports of the conditioning data are pairwise disjoint[2]. In the following algorithm, I_c denotes the subset of indices affected by conditioning (e.g. $I_c = \cup_{A\in\mathcal{A}} A$):

(i) simulate $(Y_i, i \in I_c)$ *given* $\big(N_A = n_A, A \in \mathcal{A}\big)$*;*
(ii) simulate $(Y_i, i \in I)$ *given* $(Y_i = y_i, i \in I_c)$*;*
(iii) simulate $\big(N_i, i \in I\big)$ *given* $(Y_i, i \in I)$ *and* $\big(N_A = n_A, A \in \mathcal{A}\big)$.

The first step consists of simulating the distribution

$$g_c(y_i, i \in I_c) \propto g(y_i, i \in I_c) \prod_{A\in\mathcal{A}} e^{-z_A} z_A^{n_A}$$

where $z_A = \sum_{i\in A} \varphi(y_i)$ is the potential of block A. This can usually be achieved using an acceptance-rejection method unless the number of elements $\#I_c$ of I_c is large, in which case the Gibbs sampler is required (Geman and Geman, 1984; Kleingeld *et al.*, 1997). The second step is simply a conditional simulation of a Gaussian vector. Regarding the third step, three different cases have to be considered. Each component N_i of $(N_i, i \notin I_c)$ is independently Poisson distributed with mean $z_i = \varphi(y_i)$. If $\#A = 1$, say $A = \{i\}$, then we put $N_i = n_i$. Finally, if $\#A > 1$, then the vector $(N_i, i \in A)$ is simulated according to a multinomial distribution with index n_A and parameters $(z_i/z_A, i \in A)$.

3 Application of the Cox model by examples

The south-western African coast hosts significant diamond deposits within gravels associated with paleo-shorelines, channels and transgressive lags (Kuhns, 1995). The development of the marine diamond deposits was complex and involved the interaction of fluvial, shoreline and aeolian sedimentary environments on a stable continental shelf under conditions of changing sea levels. The deep water ($> 70m$ water depth), offshore deposits comprise low-grade, aerially extensive, but thin, composite marine lag gravels (Corbett, 1996). The current mining technology requires a highly selective mine plan to facilitate profitable exploitation. Sampling costs for these offshore deposits are high and consequently, optimization of these programmes and an understanding of the confidence of the estimates are essential to ensure successful exploitation.

3.1 SAMPLE OPTIMIZATION

The objective of this study was to devise the optimum sample support size and pattern to evaluate and ultimately selectively mine a particular marine diamond deposit. At first, a stochastic model of the diamond content was created using

[2] Define new blocks by removing the indices of the samples within them, and update their numbers of particles.

geological information and reconnaissance sampling of the actual target area, as well as a regional understanding of the environment. Then, different sampling campaigns of varying sample sizes and patterns were simulated, from which block estimates and mine plan selections were calculated. Finally, a financial analysis of the various sampling campaigns and their expected mining revenues was carried out to determine the most optimum sampling strategy.

The topography of the deposit has favored the long-term preservation of gravel accumulations, which constitute semi-permanent trapsites that concentrate diamonds and other heavy minerals. The reconnaissance sampling data comprised only 63 widely spaced samples, and experience has shown that a limited dataset of this size creates a constrained tail to the overall stone density of the deposit. However, the stone distribution tail can be better modeled using information from analogous deposits which showed an overall Sichel distribution (Sichel, 1973) and a significantly higher variance (10.85) than the sample data (see Figure 1).

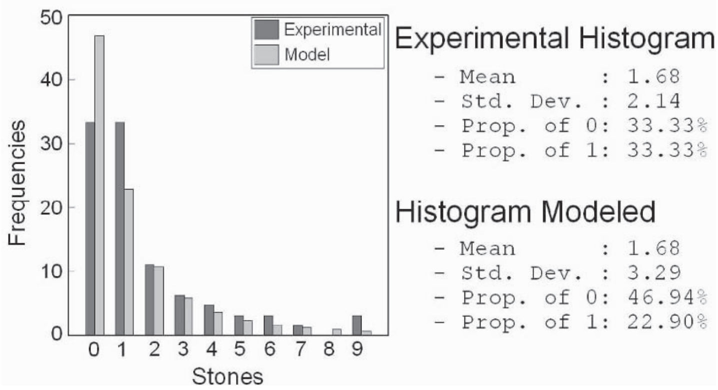


Figure 1. Modeled histogram of the stone distribution compared to the experimental stone distribution.

The variogram of the deposit was also derived from similar deposits (33% nugget effect; anisotropic spherical structure with ranges $345m$ (305°) and $145m$ (215°)). The model assumptions were validated by analyzing the statistics produced by a set of non-conditional simulations performed at the conditioning data points (see Figure 2).

Twenty conditional simulations of the diamond stone density were created over the entire target area using the Cox process. Geologists also visually reviewed the realizations and confirmed their validity against conceptual geological models. A series of sampling campaigns were then designed according to financial considerations and available sampling tools. Various sample sizes (from $12.2m^2$ to $700m^2$), sample shapes (trench, block and drill) and sample spacings ($50m \times 50m$ to $250m \times 50m$) were considered.

Ordinary kriged estimates of the stone density for $50m \times 50m$ blocks were performed on each sampling campaign and each realization. The block estimates could then be compared to their respective "actual" values in the simulated deposits

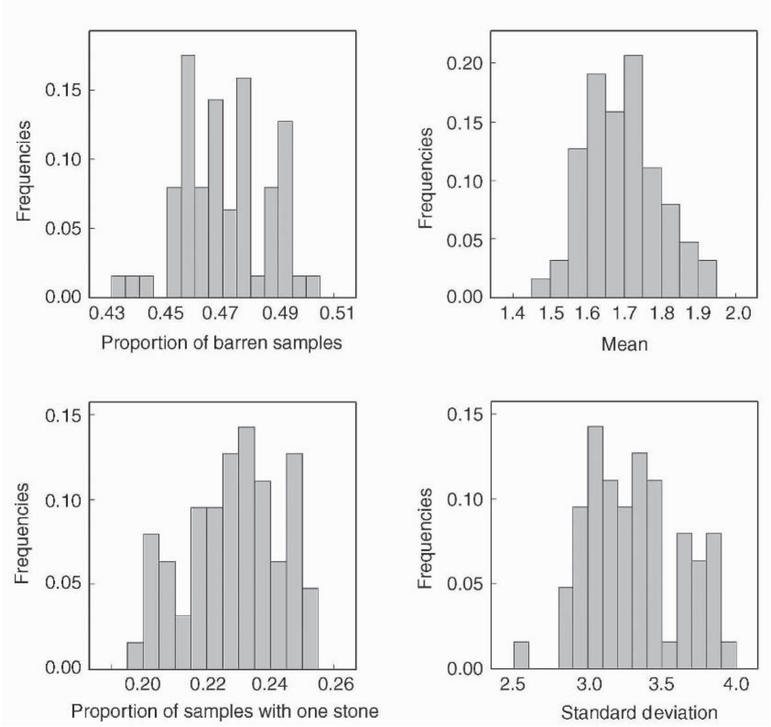


Figure 2. Histograms of the unconditional simulations at data points to validate the simulation model.

(regularized into $50m \times 50m$ blocks). Regressions comprising of the "actual" grades to the estimates for each sampling campaign per realization are summarized in Table 1.

The results of the comparison between estimated and "actual" block grades could be related to the characteristics of the source sample data. Sampling campaigns with few large samples (trench or bulk) returned accurate but highly smoothed mean estimates (a function of poorly defined spatial structure). Sampling campaigns with numerous small samples (drills) returned less accurate mean estimates, but defined the grade variability more accurately (a function of better defined spatial structure). The accuracy of the estimates relative to the "actual" grades can be shown graphically (see Figure 3).

Selective mine plans based upon the estimated grades were then set up with a realistic cut-off grade that resulted in a high degree of selectivity (only 25% of the target area is above cut-off). The actual contribution (revenue - mining cost) for each block was then calculated using the simulation as the "actual" resource grade. Thereby, the mining profits of the selective mine plans for each sampling campaign/estimation combination were established. The cost of sampling was then deducted from the mining profit and the revenue that was generated from the sampling recovery added. Finally, an overall profit was established for each

Table 1. Statistical results of the block estimates for each sampling campaign.

Type	Size $(m \times m)$	Area (m^2)	Spacing $(m \times m)$	Mean	Standard deviation	Coefficient of variation	Regression
trench	14×50	700	250×50	0.768	0.366	0.476	0.694
trench	14×50	700	150×100	0.792	0.304	0.384	0.816
trench	14×25	350	100×100	0.775	0.464	0.599	0.894
bulk	25×25	625	150×100	0.796	0.321	0.404	0.828
bulk	14×14	196	100×100	0.776	0.426	0.548	0.872
drill	3.5×3.5	12.25	100×100	0.757	0.458	0.604	0.837
drill	3.5×3.5	12.25	100×50	0.755	0.508	0.674	0.902
drill	3.5×3.5	12.25	50×50	0.768	0.616	0.802	0.945

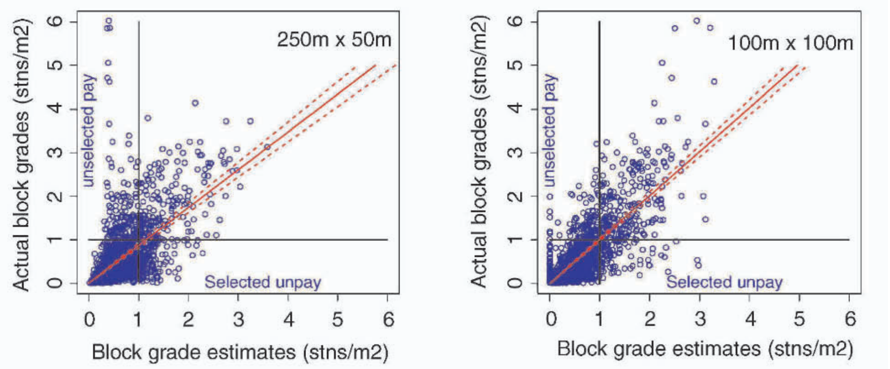


Figure 3. Scatterplots of block estimates versus ”actual” block grades for two different sampling campaign realizations.

sampling campaign. Table 2 presents the financial results obtained. It shows the percentage mining and overall (assuming perfect knowledge) profits obtained by the selective mining for each of the different sampling campaigns. It appears that the trench sampling campaigns produce less optimal returns for both mining and overall profits. The high degree of selectivity penalizes these estimates despite lower net sampling costs. The bulk samples realize slightly better returns for mining profit and a reasonable overall profit because of the associated low sampling costs. The drill sampling produces good returns from mining and overall profits, a consequence of more accurate estimation, however the high cost of drill sampling penalizes the overall profit, with the $100m \times 50m$ drill sample pattern producing the optimal result.

Table 2. Summary of the selective mining of the simulated deposit based upon various sampling campaigns.

Type	Size $(m \times m)$	Area (m^2)	Spacing $(m \times m)$	Number of samples	Mean percentage of mining profit	Mean percentage of overall profit
trench	14×50	700	250×50	231	59%	52%
trench	14×50	700	150×100	180	72%	66%
trench	14×25	350	100×100	270	82%	77%
bulk	25×25	625	150×100	180	73%	68%
bulk	14×14	196	100×100	270	76%	73%
drill	3.5×3.5	12.25	100×100	270	81%	76%
drill	3.5×3.5	12.25	100×50	567	90%	82%
drill	3.5×3.5	12.25	50×50	1080	93%	81%

3.2 CONFIDENCE LEVELS OF BLOCK ESTIMATES

The objective of this second study was to determine the relative confidence of block estimates of stone density (expressed as stones per square meter) for another marine diamond deposit. The approach adopted here consisted of creating conditional simulations of the deposit, deriving the stone density distribution of each block, and deducing confidence limits for each block estimate.

The sampling data comprised 544 large diameter drill samples ($7m$ diameter), predominantly arranged on a $50m$ square grid. A section of the deposit was less densely sampled ($70m$), whereas another portion was unsampled (see Figure 4).

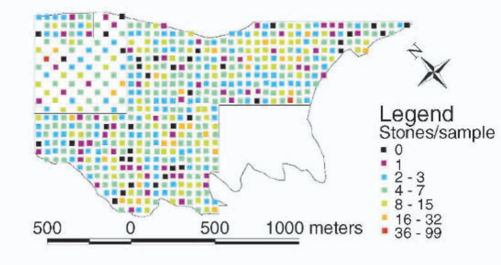


Figure 4. Locality plot showing the large diameter samples.

The stone distribution was modeled as a Sichel distribution using the experimental mean and the proportion of barren samples. The sample variogram of the deposit was determined from the sample data and inferences from regional geological structures (34% nugget effect; two isotropic nested spherical structures, 50% of the sill at $76m$ and 16% at $300m$), from which a variogram in Gaussian space was derived. The Cox model was validated by analyzing the statistics produced by a set of unconditional simulations at the locations of the conditioning data points.

One hundred and twenty-five conditional Cox simulations were then created at the sample support size and averaged into $50m \times 50m$ blocks. An example of a realization at the sample support size, together with the conditioning sample data is presented in Figure 5.

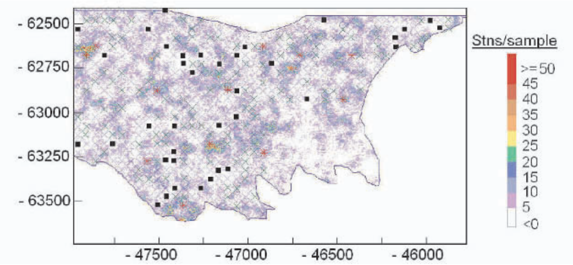


Figure 5. Conditional simulation of a marine diamond deposit.

From the simulations, statistics for each block were calculated and compared with ordinary kriged estimates, using the same dataset and the same variogram model (see Figure 6).

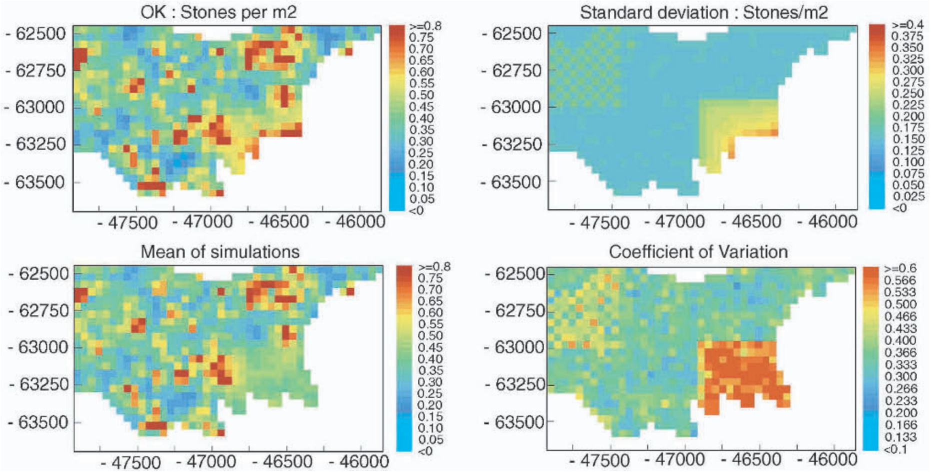


Figure 6. Comparison between kriged and simulated results.

The kriged estimate (top left) and the mean of the simulations (bottom left) of the stone density show good similarity with compatible low and high grade areas. The estimation variance (top right) provides an indication of the data density for these estimates as this parameter highlights areas of similar sample spacings. The coefficient of variation, measuring the relative variability of the distribution of possible grades, provides further information by incorporating the relevant sample grade information. The simulated distribution of each block is also used to guide mining selections when a cut-off grade is considered. Figure 7 shows a comparison between the blocks whose kriging estimate lies above a given cut-off grade, and those whose simulated grade has more than 65% chance of lying above the same cut-off grade.

In this case study, the simulation based selection, using the probability above cut-off selection criteria, selected 56% of those blocks selected above cut-off of the kriged estimates. However, the reduced ”probability” selection contains 89% of the total diamonds. More generally, this approach allows the relative risk of the mining selection to be incorporated into mine-planning decisions and has proved to return better results than a single grade cut-off method.

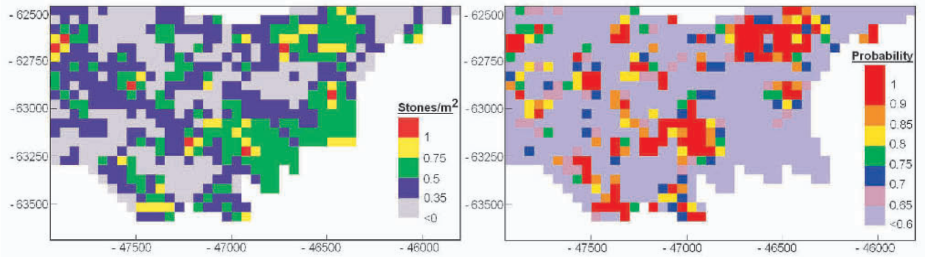


Figure 7. Mining selections above a cut-off grade based on kriged estimates (left) and 65% or higher probability of above a cut-off grade based on Cox simulations (right).

4 Conclusion

The development of the Cox process to model the spatial distribution of diamonds within marine deposits has enabled the successful optimization of sampling campaigns in terms of both sample size and pattern. The quantification of block grade uncertainty from conditional simulations has also proven to significantly assist mine planning decisions.

References

Corbett, I.B., *A Review of Diamondiferous Marine Deposits of Western Southern Africa*, Africa Geoscience Review, Vol. 3-2, 1996, pp. 157-174.

Feller, W., *An Introduction to Probability. Theory and Applications Vol. 2*, Wiley, New York, 1971.

Geman, S. and Geman, D., *Stochastic Relaxation, Gibbs Distribution and the Bayesian Restoration of Images*, IEEE Trans. Pattern Anal. and Mach. Intel., Vol. 6, 1984, pp. 721-741.

Kleingeld, W.J., *La géostatistique pour des Variables Discrètes*, Ph.D. Thesis, School of Mines of Paris, 1987.

Kleingeld, W.J., Thurston, M.L., Prins, C.F. and Lantuéjoul, C., *The Conditional Simulation of A Cox Process with Application to Deposits with Discrete Particles*, In Baafi E.Y. and Schofield N.A. (eds.), Geostatistics Wollongong ’96, Vol. 2, Kluwer, Dordrecht, 1997, pp. 683-694.

Kuhns, R., *Sedimentological and Geomorphological Environment of the South African Continental Shelf and its Control on Distribution of Alluvial, Fluvial and Marine Diamonds*, Society for Mining, Metallurgy and Exploration, Proceedings of Annual Meeting, Denver , 1995.

Sichel, H., *Statistical Valuation of Diamondiferous Deposits*, In Salamon M.D.G. and Lancaster F.H.(eds.), Application of Computer Methods in the Mineral Industry, The South African Institute of Mining and Metallurgy, Johannesburg, 1973, pp. 17-25.

INVERSE CONDITIONAL SIMULATION OF RELATIVE PERMEA- BILITIES

JAIME GÓMEZ-HERNÁNDEZ
Universidad Politécnica de Valencia, Spain
CAROLINA GUARDIOLA-ALBERT
Universidad Politécnica de Valencia, Spain
currently with Instituto Geológico y Minero de España, Madrid, Spain

Abstract. The spatial variability of relative permeability curves has not attracted much attention yet. This paper addresses this issue, and extends the self-calibration technique for the generation of absolute and relative permeabilities conditioned not only to permeability data but also to saturation and pressure data.

The paper starts with a sensitivity analysis presenting a synthetic example where the spatial variability of relative permeability is relevant, then proceeds with a derivation of the algorithm used to condition a realization of relative permeabilities to pressure and saturation data (both steady state and transient) and concludes with the demonstration of the technique with one synthetic example.

The paper shows that the spatial variability of relative permeabilities is important in reservoir characterization. It also demonstrates how the self-calibrating simulation method can be used to generate realizations of spatially variable relative permeability curve parameters which are consistent with measured values of the state of the reservoir.

1 Introduction

Stochastic modeling of multi-phase flow in heterogeneous porous media is becoming common practice in petroleum engineering and subsurface hydrology. Inverse modeling theory provides a methodology to integrate both static and dynamic data in reservoir characterization. Absolute permeability is one of the parameters that are typically estimated with inverse conditional or unconditional simulations, whereas relative permeabilities are assumed to be known homogeneous functions within the reservoir. However, when studying multiphase flow, relative permeability is the parameter that controls the rate of displacement of the different phases present in the reservoir. This paper discusses the characterization of the spatial variability of relative permeabilities by inverse conditional simulation. A new inverse technique to estimate spatial distributions of both absolute and relative permeability parameters has been developed based on the self-calibrating method. Since relative permeabilities are dependent on one of the state variables

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 325-331.

(saturation), the optimization problem is highly non linear, increasing the difficulty of the inverse problem that has to be solved.

2 Sensitivity Analysis

The governing equations for immiscible two-phase flow are formulated in terms of water saturation and fluid pressure. Substituting a generalized form of Darcy's law into the mass conservation equation, and neglecting gravity effects, the diffusivity equation for the horizontal flow of each fluid is obtained:

$$\frac{\partial\left(\phi\rho_l S_l\right)}{\partial t}-\nabla\cdot\left[\rho_l\frac{\mathbf{k}k_{rl}}{\mu_l}\nabla p_l\right]=-q_l\quad\text{for } l=w,o \tag{1}$$

where subscripts w and o refer, respectively, to water (wetting phase) and oil (non-wetting phase), $\mathbf{k}$ is the absolute permeability tensor $[L^2]$, k_{rl} the dimensionless relative permeability for phase l, S_l is the saturation, ϕ is the porosity (dimensionless parameter), ρ_l is the fluid density $[M/L^3]$, μ_l is the viscosity $[ML/T]$, p_l is the fluid pressure $[M/LT^2]$, q_l is the injection or production rate per unit volume $[T^{-1}]$ and t is time $[T]$.

The above equations are solved by finite differences after neglecting gravity and capillary pressure terms and assuming the following constitutive equation relating relative permeability and saturation (Brooks and Corey, 1966):

$$k_{rl}=k_{rl}^0\left(\frac{S_l-S_{rl}}{1-S_{rl}-S_{rl'}}\right)^{n_l}\quad\text{for } l=w,\ l'=o \text{ or } l=o,\ l'=w \tag{2}$$

in which n_w, n_o, S_{rw}, S_{ro}, k_{rw}^0 and k_{ro}^0 are parameters defining the relationship. For the purpose of characterizing the heterogeneity of relative permeability, we will assume that exponents n_w and n_o are homogeneous, but that the remaining coefficients, i.e., the residual saturations for oil and water (S_{rw}, S_{ro}), and the end-points of the relative permeability functions may vary within the reservoir.

To demonstrate the importance of accounting for the spatial variability of relative permeabilities, two different runs are performed, both of them with the same heterogeneous absolute permeability field, but with different relative permeability fields. The first run is done assuming homogeneous relative permeabilities, the second run is done with heterogeneous values. Conditional simulations are used to construct absolute and relative permeability fields. The absolute permeability fields were generated by GCOSIM3D (Gómez-Hernández and Journel, 1993), following a lognormal distribution with known mean and variance, and a spherical variogram. The relative permeability fields are constructed by generating each of the four parameters defining relation (2) as a Gaussian field with known mean and variance and a Gaussian variogram (it is assumed that these parameters vary smoothly in space).

The 2D spatial domain mimics a quarter of a five-spot, discretized in 15 × 15 grid blocks of size 10 m ×10 m. The injector well is located at the left lower corner, and the production well at the right upper corner. The initial conditions for the 2D run are $S_w = 0.1$ and $p = 6.895 \cdot 10^6$ Pa. Water is injected at a constant rate of 2.5 kg/s. Constant 4.8-hour time steps are prescribed, and the simulations are run for 120 days.

For the first run, a single relative permeability curve for each fluid is used, with parameter values equal to the mean values used for the generation of the heterogeneous field: $k_{rw}^0 = 0.7$, $S_{rw} = 0.1$, $k_{ro}^0 = 0.85$ and $S_{ro} = 0.2$. The second run uses heterogeneous values for all four parameters, maintaining the same mean values as the previous run. The saturation and pressure fields at the end of the 120 days are shown in Figures 1 and 2.

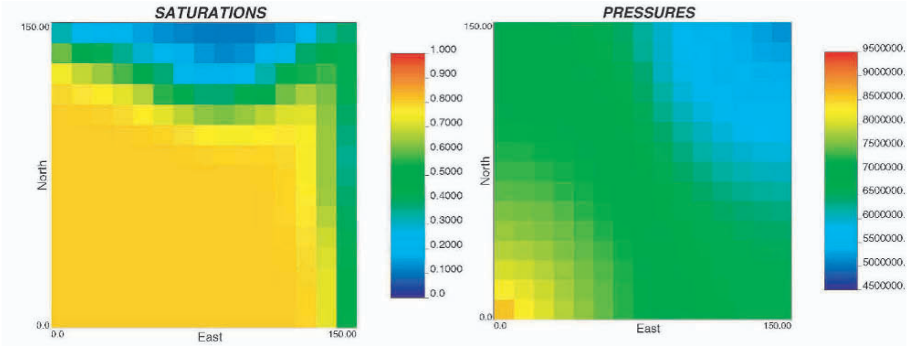


Figure 1. Saturation and pressure front at $t = 120$ days. Absolute permeability is heterogeneous, but relative permeability curves are homogeneous. The water injection well is located at the lower left grid block and the production well at the upper right grid block.

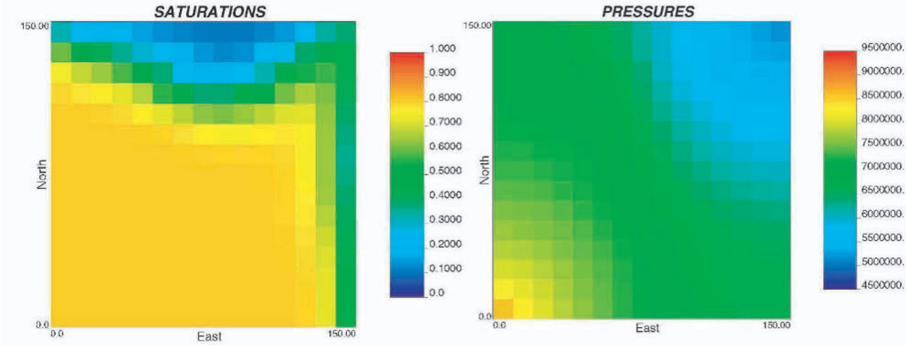


Figure 2. Saturation and pressure front at $t = 120$ days. Absolute and relative permeability are heterogeneous. Same absolute permeabilities and well configuration as in previous figure.

From these figures it can be observed that pressures are not very much affected by the heterogeneity of the relative permeability curves, but that saturations are. Similar conclusions were obtained in 1D by Guardiola-Albert and Gómez-Hernández (2002).

3 Inverse conditional simulation algorithm

The optimization algorithm is based on gradient methods, and the concept of master points is borrowed from the Sequential Self-Calibrated method (Gómez-Hernández et al., 1997; Hendricks-Franssen, 2001). Calibration of the flow model to non linearly related data is formulated as an optimization problem, which tries to minimize an objective function. A computer code was written to couple the forward two-phase flow simulator TOUGH (Pruess and Oldenburg, 1999) with an iterative inverse method. After calibration, the result is a plausible representation of the reservoir honoring historical pressure and saturation data. Calibration parameters are: absolute permeabilities, the two end-points of oil and water relative permeability functions, and the two residual saturations. For the sake of simplicity, the shape parameters ($n_o and n_w$) are constant and equal to 1. The main steps in the iterative process, followed by the inversion technique developed, are summarized here. The loop from step 2 to step 7 is repeated until convergence is reached.

For each iteration ^{IT}:

1. Generate a conditional or an unconditional simulation of the four parameters k^0_{rw}, k^0_{ro}, S_{rw} and S_{ro} and of the absolute permeability, k. The generated fields constitute the seed or initial input fields.
2. The two-phase flow numerical solver is run. Saturation and pressure fields are obtained for all the time steps at every grid block.
3. Evaluate the following objective function:

$$J^{IT} = \sum_{t=1}^{T_s}\sum_{i=1}^{N_s} w_{s,i}\left(S_{i,t}^{SIM,IT} - S_{i,t}^{MEAS}\right)^2 + \sum_{t=1}^{T_p}\sum_{i=1}^{N_p} w_{p,i}\left(p_{i,t}^{SIM,IT} - p_{i,t}^{MEAS}\right)^2 \tag{3}$$

   where N_s and N_p are the number of saturation and pressure data points, respectively, T_s and T_p are the number of times at which saturation and pressure have been measured. Indices ^{SIM} and ^{MEAS} indicate simulated and measured values, and, w_s and w_p are weighting factors.
4. If J is smaller than a pre-determined value, the simulated permeability values are said to be conditioned to the measured saturation and pressure values, and the iterative loop stops. On the contrary, the optimization continues and the k, k_{rw}, k_{ro}, S_{rw}, and S_{ro} fields are perturbed.
5. The optimization procedure determines the value of the perturbation that is applied to the initial field so that the objective function is reduced.
6. Go to step 2. The modified reservoir model is input again into the reservoir simulator.

4 Synthetic example

To check the feasibility of the inverse technique a simple synthetic example is presented. Absolute permeability is heterogeneous all over the reservoir, and relative permeability is piecewise heterogeneous as shown in Figure 3.

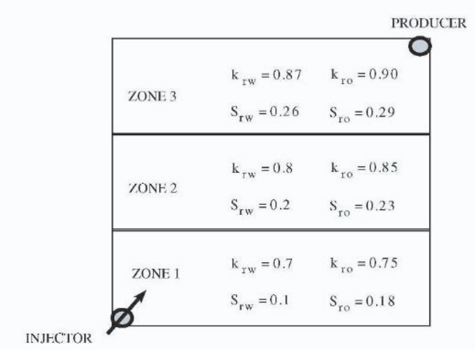


Figure 3. The reservoir is divided into 3 different zones, within each of them the relative permeability parameters are assumed to be constant. The reference values for the relative permeability parameters are given in the figure.

No statistical correlation was considered between the five parameters (k, k^0_{rw}, S_{rw}, k^0_{ro} and S_{ro}). However, there is an implicit correlation because all the parameters are calibrated to the same set of production data. A reference run as performed with the values shown in Figure 3, which was sampled at five locations and at ten time steps to be used as the conditioning information for the inverse conditional simulation.

Scatterplots for saturation and pressure at well locations for all sampling times are shown in Figure 4. They compare the degree of mismatch between the seed fields and the calibrated fields to the data (on the left, the seed fields, on the right the calibrated fields). The calibration reduces considerably the spread of the scatterplots, reflecting the effect of jointly conditioning absolute and relative permeabilities to the available saturation and pressure data. We could consider the calibrated field as a plausible representation of a reservoir for which only partial historical evolution is known.

5 Conclusions

A code has been developed and implemented for the simultaneous generation of absolute and relative permeability fields conditioned to historical production data. The heterogeneity of relative permeability is described by the heterogeneity of the parameters that describe a specific relationship between saturation and relative permeability.

Four parameters in this relationship (two end-points of the curves and the residual saturations) were chosen to characterize the relative permeability curves. Joint conditioning of absolute and relative permeabilities to pressure and saturation data considerably improved the history match in the synthetic example presented.

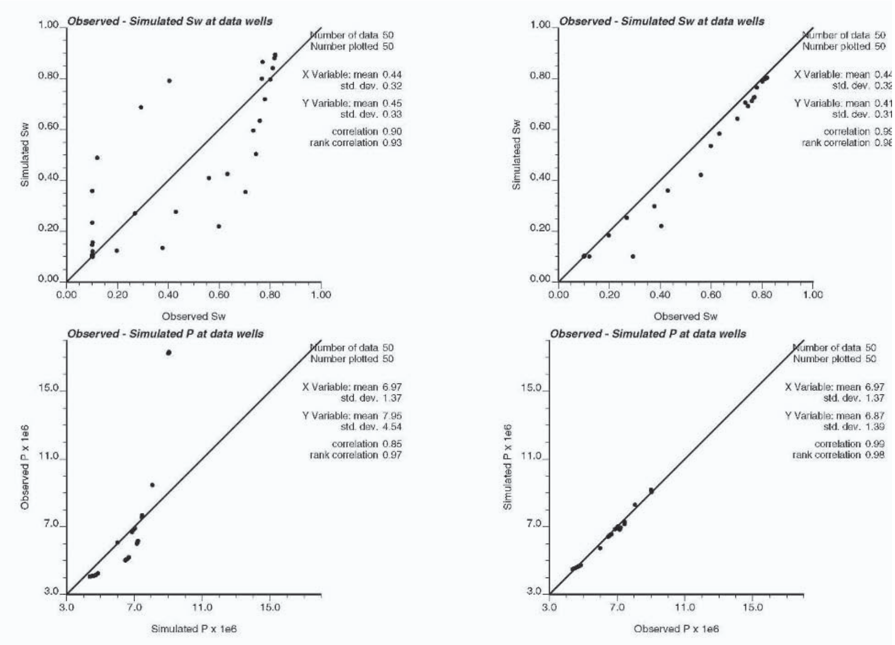


Figure 4. Simulated versus "observed" reference values are plotted before (left) and after (right) the inversion is performed. Upper scatterplots represent water saturations and lower scatter plots pressures.

Acknowledgements

The second author is grateful to the fellowship by the Repsol Foundation that financed her stay at Universidad Politécnica de Valencia during her Ph.D. studies. Partial funding from the Spanish National R&D Plan, through project REN2002-02428, and from the Spanish Nuclear Waste Management Company, ENRESA, through R&D contract 774325 are acknowledged.

References

J.J. Gómez-Hernández and A.G. Journel, *Joint simulation of multiGaussian random variables*, Oxford University Press, 1993.

R.H. Brooks and A.T. Corey, *Properties of porous media affecting fluid flow*, J. Irrigation and Drainage Division. Proceedings of the American Society of Civil Engineers, 1966.

Hendricks-Franssen, H. J., *Inverse Stochastic Modelling of Groundwater Flow and Mass Transport*, PhD dissertation. Universidad Politécnica de Valencia, Spain, 2001.

J. Gómez-Hernández and A. Sahuquillo and J.E. Capilla, *Stochastic Simulation of Transmissivity Fields Conditional to Both Transmissivity and Piezometric Data. 1. Theory*, Journal of Hydrology, vol. 203, 1997, p. 162-174.

C. Guardiola-Albert and J. Gómez-Hernández, *Inverse modelling of two-phase flow: calibration of relative permeability curves*, Calibration of Reliability in Groundwater Modelling: A Few

Steps Closer to Reality (Proceedings of ModelCARE'2002, Prague, Czech Republic. IAHS, Publ. no. 277, 2002, p. 190-195.

K. Pruess and C. Oldenburg. Tough2 user 's guide. Report LBNL-43134, Ernert Orlando Lawrence Berkeley National Laboratory, Berkeley, 1999.

This page intentionally blank

QUANTIFIABLE MINERAL RESOURCE CLASSIFICATION: A LOGICAL APPROACH

CHRISTINA DOHM
Mineral Resource Evaluation Department (MinRED), Exploration Division of AAplc, Johannesburg, South Africa

Abstract. In terms of the reporting codes Mineral Resource classification is a function of increasing confidence in the geoscientific information and the associated resource estimate. An overview of Mineral Resource classification approaches is given; the tendency in resource classification is to concentrate on the confidence associated with the grade estimate. Uncertainties linked to tonnage and metal estimates are rarely explicitly mentioned. As for the risk associated with the underlying geological model it is often, if at all, only considered on a global rather than a local basis. The objective is to present a quantifiable Mineral Resource classification guideline that recognises uncertainty in both geological and resource models, considers confidence in estimation of metal content for specified production periods and also takes into account both the correlation of blocks in the block model as well the change of support between an estimated block and the production period. This classification method builds on a previous publication (Dohm, 2003), where a technique for assessing the combined risk associated with both the geological and grade models was demonstrated. The final result is a succinctly classified mineral resource model, which is based on objective quantifiable classification rules that recognises the uncertainty related to subjective interpretations of the available information.

1 Introduction

The classification of Mineral Resources and Ore Reserves forms an integral part of Mineral Resource evaluation and reporting. Mineral Resource classification categories correspond to an increasing function of geoscientific knowledge and confidence. A Mineral Resource is classified as Inferred if the tonnage, grade and mineral content can be estimated with low confidence. Indicated Mineral Resources represent that part of the Mineral Resource for which tonnage, densities, shape, physical characteristics, grade and mineral content can be estimated with a reasonable level of confidence. For Measured Mineral Resources these attributes can be estimated with a high level of confidence. Only Measured and Indicated Mineral Resources can be converted to Ore Reserves. In terms of the guidelines of the reporting codes the Competent Person (JORC, SAMREC) or Qualified Person (NI 43-101) is to provide a view of the relative confidence the investment community should place on the published Mineral Resources and Ore Reserves of mining and exploration companies.

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 333-342.

The main elements that affect the confidence in the resource estimate are the reliability of the geological model, the continuity of the mineralisation, the sampling grid configuration, the quality of the sampling data, and the reliability of the evaluation method. The most important element is the interpretation of the geology and the delineation of the resource (Stephenson, 2001). In practice the level of uncertainty in the geological model is often not easily incorporated in the Mineral Resource classification. In many cases global discount factors are applied to take account of the unpredictability of the geological features.

A number of approaches that the author encountered during project reviews presented here illustrate the evolution of Mineral Resource classification methodologies, concluding with a holistic Mineral Resource classification guideline. This guideline recognises uncertainty in both geological and resource models, considers confidence in estimation of tonnage, grade and metal content for specified production periods and also takes into account both the correlation of blocks in the block model as well the change of support between an estimated block and the production period.

2 Questionable Mineral Resource classification strategies

Since the Bre-X scandal the spotlight has been focussed on Mineral Resource classification methodologies. Two interesting but not recommended classification strategies observed during project appraisals in recent years are discussed.

2.1 NUMBER OF SAMPLES PER BLOCK OR PER UNIT AREA

The crudest set of classification rules the author has come across is: One drillhole per hectare identifies an Indicated Mineral Resource, more drillholes per hectare allow for the resource to be classified as Measured and when there are no drillholes but the area is within the mining lease it can be considered as an Inferred Mineral Resource.

This method does not take cognisance of the spatial continuity of the mineralisation, and anisotropy, if it should exist, is also ignored. Change of support is also not considered as 10000 square metres (1ha) can be achieved in a number of ways for example a rectangle (40m x 250m) or square (100m x 100m) are equivalent in this scheme.

The relative locations of the drillholes are not recognised; for example four 2x2, 1ha squares with a set of clustered drillholes close to the four touching corners will be considered Indicated as will four 1ha squares, with evenly spaced drillholes at their centroids, be classified as Indicated.

2.2 RESOURCES WITHIN A PRODUCTION PLANNING PERIOD OF RESERVES

In this two-dimensional example ordinary kriged estimates were produced for the entire mine lease area. Blocks were assigned the average grade of the deposit when their estimation became “unreliable”, e.g. the criterion for the minimum number of samples in the search volume is not met. The classification rule applied here was established from time-based production planning considerations. Resources were classified as a consequence of the reserve classification and not the other way round. The argument put forward was that grade control information acquired during mining activities would be adequate to support the classification.

Planned 5-year and 20-year production period mining outlines were used to differentiate between Proved and Probable Ore Reserves. The Proved Ore Reserves incorporated the Measured Mineral Resources and the Probable Ore Reserves partially incorporated the Indicated Mineral Resources. Resources lying beyond the 20 year planning limit were also defined as Indicated if sufficient drilling information was available. The area between the 20-year plan and the lease boundary were to be considered as Probable Ore Reserves. Inferred Mineral Resources were non-existent. It is obvious that this set of classification rules was unacceptable and required revision. On recommendation the company adopted a quantifiable risk based classification strategy, still related to production periods but independent of mining lease boundaries and also including the confidence of the resource estimates.

3 Range of influence of the variogram model and Resource Classification

Methods in place for classifying resources are often based on the kriging variances of grade estimates or functions of the variogram parameters and kriging parameters. The semi-variogram of a mineral deposit reflects the spatial variability of the sample grades at fixed distances and along a given direction. Snowden (1996) suggests interpreting this spatial continuity to determine appropriate drillhole patterns to achieve various levels of confidence in resource classification.

Resources are classified as Inferred when drillholes are further apart than the range of influence of the variogram. The drill spacing at which a distinction between Measured and Indicated is made is based on a rule of thumb and is taken as the distance equivalent to two thirds of the total variability i.e. two thirds of the sill of the variogram model.

The ranges of the variogram are not sufficient for resource classification; the nugget effect will for instance have a significant influence in this classification. If the nugget effect is high and the structured component is relatively short, as is common for Witwatersrand gold deposits, this classification method will be of little use; the majority of the mineral resources will be classified as being Inferred resources.

4 Kriging variance and variations thereof in the classification scheme

One of the advantages of kriging estimation techniques is that when correctly applied these techniques produce unbiased block estimates and ensure minimum estimation variance known as the kriging variance. The kriging variance is dependent on the variogram model, and the sampling grid configuration in relation to the block that is being estimated. It is possible to calculate the kriging variance without producing the estimate. It is thus not surprising that a number of classification schemes in the past were based on the kriging variance: a few applications are given below:

4.1 INTERPOLATION, EXTRAPOLATION AND RESOURCE CONFIDENCE

This classification rule was based on the following reasoning related to the type of estimation. Measured Mineral Resources arise from interpolated blocks, which have lower kriging variances and therefore higher confidence associated with them. Indicated Mineral Resources occur when blocks are extrapolated, this means that these blocks have higher kriging variances and thus a reduced confidence is associated with them. Any blocks extrapolated beyond the range of influence of the variogram are classified as Inferred blocks.

4.2 SAMPLE VARIANCE, KRIGING VARIANCE AND NUMBER OF SAMPLES

The Mineral Resources are classified as Measured if the kriging variance of the block is less than the sample variance. If not Measured and at least 4 samples were within the maximum range of influence of the block the resource is classified as Indicated. Inferred Mineral Resources are those blocks that did not fall into the previous two categories but with a kriging variance equivalent to that of blocks in the Indicated category.

4.3 SAMPLE VARIANCE, BLOCK VARIANCE AND KRIGING VARIANCE

Blocks are classified as Measured if their kriging variance was less than the block variance. Blocks classified as Indicated have a kriging variance less than the sample variance but greater than the block variance. Blocks with an estimated mineralised proportion less than 20 % were considered to be in the Inferred category.

5 Resource classification and relative variances

Using the kriging variance, as the only measure to quantify uncertainty in block grade estimate, is questionable as the only relationship the kriging variance has to the local sample grade values is through the variogram model, which is on a global average basis rather than a local basis. This means that the kriging variance for a specific sample to block configuration is a fixed value irrespective of whether the grade values are highly variable or more uniform. It is clear that there is greater confidence in the estimate of the latter block than that of the former block. This anomaly led to the introduction of classification techniques that concentrated on relative variances that recognise the local data configuration and variability.

5.1 RELATIVE KRIGING ERRORS AND THE NUMBER OF SAMPLES

Blackwell (1998) presented an argument for introducing the Relative Kriging Variance (RKV); the ratio of the kriging variance and the kriged estimate squared. From this the Relative Kriging Standard Deviation (RKSD) is defined as the square root of the RKV. The RKSD is plotted against the number of samples used in the kriging of the block. Two threshold values for the RKSD are selected arbitrarily, but based on experience, to separate the Measured, Indicated and Inferred categories.

5.2 RELATIVE VARIABILITY INDEX

The implementation of the Relative Variability Index (RVI) as a measure of confidence in the estimate was proposed (Arik, 1999). The RVI is the ratio of the square root of the combined kriging variance and the kriged estimate of the block. The combined kriging variance is the square root of the product of the kriging variance and the local weighted average variance. The histogram of the RVI is analysed to determine thresholds for distinguishing between categories, and the proposal is to use the 50th and 90th percentile RVI values to identify the three different resource categories.

5.3 INTERPOLATION VARIANCE

Yamamoto (2000) introduced the Interpolation Variance (IV) as an alternative measure of the reliability of ordinary kriging estimates. The IV reflects the local variability as expressed by the data. It is the weighted average of squared differences between the data values and the block estimate. An advantage is that this variance recognises the proportional effect when present. This interpretation of the variance is however, only valid if and only if all the ordinary kriging weights are positive.

6 Mineral Resource classification linked to a production period

The philosophy of applying a classification rule that considers *"the % error in the estimate of the block being classified is within 15% with 90% confidence for a specific production period"* has been around for a number of years, at least since the early 1990's. This is an empirical rule that has been accepted in the mining industry. The specific production period should define the resource category: the shorter the production period, the higher the confidence category, for a longer production period the resource category is lower. Many variants of this rule are applied in practice.

The percentage error in the estimate of the mean is given by

$$\%\,\text{error} = \frac{\text{Standard Error}}{\text{Estimate}} \times 100\% = \frac{{}^{s}\!/\!{}_{\sqrt{n}}}{\bar{z}} \times 100\% = \frac{s}{\bar{z}} \times \frac{1}{\sqrt{n}} \times 100\% = {}^{\text{CoV}}\!\!\diagup\!{}_{\sqrt{n}} \times 100\%$$

CoV is the coefficient of variation; s is the standard deviation and $\bar{z}$ is the average of the samples.

Relative 90% confidence limits can be established from the product of the CoV/√n and 1.645, the standard normal deviate.

If the resource blocks can be considered independent the relative variability of a block can be converted to the equivalent of the variability in a production period by dividing the coefficient of variation of the block in the resource model by √n. Where *n* would be the number of independent blocks that would be required to represent the production period as shown below.

$$\mathrm{CoV}_{\text{Production Period}} = \frac{\mathrm{COV}_{\text{Block}}}{\sqrt{n}}$$

The estimated resource blocks are correlated and the support of the blocks are relatively small compared to the support of for instance the annual production. In general, if the histogram of the sample support is skew and if the block support is small the histogram of estimates will be skew and as the support increases the histogram of the estimated blocks in the resource model will approach normality as per the central limit theorem. This means that the histograms of estimates of production periods, which consist of many resource blocks, are expected to approach normality and independence.

Mining, though, does not take place in independent blocks thus the effect of correlation must be brought into account when an individual estimated resource block is being classified in terms of production periods.

6.1 THE "INDEPENDENT" √n

It is necessary to modify the 90% limits of the estimated blocks in the block model to represent the equivalent variability of a production period. Therefore, a factor that takes this correlation and the production period into account has to be determined to replace the "independent" square root of "n".

$$\mathrm{CoV}_{\text{Production Period}} = \frac{\mathrm{CoV}_{\text{Block}}}{\sqrt{n}} \sim \frac{\mathrm{CoV}_{\text{Block}}}{\text{Factor Production Period}}$$

Thus

$$\text{Factor Production Period} = \frac{\mathrm{CoV}_{\text{Block}}}{\mathrm{CoV}_{\text{Production Period}}}$$

The problem is to find estimates for a representative CoV of blocks in the block model and for the CoV of the production periods. As the resource has not yet been classified the CoV of the production periods cannot be established. It is nonetheless possible to determine estimates for these values from many realisations of a conditional simulation exercise.

6.2 CONDITIONAL SIMULATION AND UNCERTAINTY ASSESSMENT

In an endeavour to attain quantifiable Mineral Resource classifications, the trend in the mining industry has been towards the application of conditional simulation techniques. A specific set of drilling or sampling results provides one view of the resource, a different set will provide a different view, the luxury of a second campaign is however not always available. A fairly quick and inexpensive method for obtaining a spectrum of possible views of the global statistical and spatial characteristics of the orebody can be obtained through conditional simulation. The variability in realisations of the simulations can be interpreted to assess the uncertainty in the resource estimates.

It is assumed that conditional simulations are carried out to reflect both the uncertainty in the geological model as well as the uncertainty associated with the grade model. Dohm (2003) introduced a technique to combine conditional indicator simulations for geology and conditional sequential Gaussian simulations for grade to assess the combined uncertainty of the geological interpretation and the grade estimation.

It is, however, vital to realise that if this tool is applied to assess the risk associated with the Mineral Resource then it is important to establish the integrity of the simulation results. For example the number of simulations to be considered for assessment and validations of the reproducibility of both variogram model and histogram of the conditioning data are crucial. When these validations are not carried out, the results can lead to incorrect Mineral Resource classifications that could have disastrous effects on Ore Reserve classifications and investment decisions.

The task at hand is to establish from the conditional simulations, what the expected coefficient of variation for the estimated grade, tonnage and metal content of a real month's or year's mining would be.

6.3 THE 15% RULE – A LOGICAL APPROACH

The purpose is to produce a measure of confidence in the resource estimate. The specific classification rule considered here is based on two production periods, namely a monthly production period for Measured Mineral Resources and an annual production period for Indicated Mineral Resources.

Critical to this resource classification guideline are the following three CoV values:

$\text{CoV}_{\textbf{Local}}$: a typical CoV for blocks of the same size as used in the estimation model.

$\text{CoV}_{\textbf{Monthly}}$: CoV signifying the relative variability of a monthly production period.

$\text{CoV}_{\textbf{Annual}}$: CoV signifying the relative variability of an annual production period.

To establish the monthly and annual CoV values the moving block technique is applied. Every realisation of the conditional simulation is "cookie cut" by units representing likely 'production periods' at various positions within the orebody. This process is repeated for a sufficiently large number of times, e.g. at least 120 months (10years) per realisation of the simulation exercise.

Representative CoV values for monthly and annual production periods can be calculated from the statistical analysis of these two supports.

The proposed method has been applied to a Zn deposit. The orebody comprises two superimposed mineralised horizons, which are structurally controlled and are part of an overturned fold limb. Both orebodies comprise a well mineralised massive sulphide horizon close to the footwall and an overlying iron formation containing banded and disseminated sulphides. Fan drilling is performed from hanging- or footwall drives, on 20 m spaced north–south sections. Intersection spacing on section varies between 10 m and 40 m depending on the complexity of the orebody. Deeper exploration drilling is on a 100m x 50m grid, sampled every 2m.

The three critical CoV values for classification obtained from 40 conditional simulations of the %ZN values, carried out in unfolded space were:

$\mathrm{CoV}_{\mathbf{Local}} = 0.754$ the relatively large CoV for a block in the resource model confirms the earlier remark that blocks are correlated.

$\mathrm{CoV}_{\mathbf{Monthly}} = 0.1564$ The CoV of monthly periods is less than that of the block.

$\mathrm{CoV}_{\mathbf{Annual}} = 0.0667$ The CoV of the annual production is as expected significantly less

The monthly adjustment factor is then calculated from

$$\text{Factor Monthly Production Period} = \frac{\mathrm{COV}_{\text{Block}}}{\mathrm{COV}_{\text{Monthly}}} = \frac{0.754}{0.1564} = 4.821$$

The annual adjustment factor is then calculated from

$$\text{Factor Annual Production Period} = \frac{\mathrm{COV}_{\text{Block}}}{\mathrm{COV}_{\text{Annual}}} = \frac{0.754}{0.0667} = 11.304$$

For a Measured Mineral Resource where the error in the monthly production estimate has to be within 15% with 90% confidence the threshold value is:

$$\mathrm{CoV}_{\text{Measured}} = \frac{0.15 \times 4.821}{1.645} = 0.440$$

For an Indicated Mineral Resource where the error in the annual production estimate has to be within 15% with 90% confidence the threshold value is:

$$\mathrm{CoV}_{\text{Indicated}} = \frac{0.15 \times 11.304}{1.645} = 1.030$$

Each block in the estimation model is considered in turn and its coefficient of variance, CoV $_{\textbf{block estimate}}$ is calculated

$$\text{CoV}_{\text{block estimate}} = \frac{\sigma_K}{Z_K}$$

Where $Z_{\mathbf{K}}$ is the kriged estimate and $\boldsymbol{\sigma}_{\mathbf{K}}$ is the kriging standard deviation of the block.

The CoV $_{\textbf{block estimate}}$ value of each block in the estimation model is then compared to the above threshold values and the decision rules are:

If the CoV $_{\textbf{block estimate}} \leq$ CoV $_{\textbf{Measured}}$ then the block is classified as Measured.

If CoV $_{\textbf{Measured}} <$ CoV $_{\textbf{block estimate}} \leq$ CoV $_{\textbf{Indicated}}$ then the block is classified as Indicated.

If the CoV $_{\textbf{block estimate}} >$ CoV $_{\textbf{Indicated}}$ then the block is classified as Inferred.

Once all the blocks in the estimation model have been classified the Measured and Indicated Mineral Resources can be considered for conversion to Proved and Probable Ore Reserves. It is recommended, that as with any automated mathematical process, the classified Mineral Resource model be validated; at least visually.

The final result is a succinctly classified Mineral Resource model, which is based on objective quantifiable classification measures that recognise the uncertainty related to subjective interpretations of the available information.

7 Comments and discussion

It is essential to ensure the integrity of the simulations by validating the inherent and spatial; variability of each realisation in terms of the histogram and variogram of the conditioning data.

An advantage of applying the conditional simulation techniques is that once the resources have been converted to reserves it is possible to compare the variability of the actual mine plan with that expected from the simulations.

The author has come across two other approaches for determining “n”, the number of independent blocks to use in the above classification. The first method determines “n” as the number of independent production blocks required to reach the range of the variogram. In the second method “n” is calculated as the ratio of the production period tonnage divided by the block tonnage. In both cases the square root of “n” is used as the divisor for the CoV to determine the 90% confidence limits.

The classification guideline proposed does not assume that the variance reduction factor should be in terms of a square root and uses CoV measures to address the change of support effect.

8 Conclusion

A number of Mineral Resource classification methodologies were discussed showing the development of understanding and incorporating uncertainty associated with the grade estimates. The final classification guideline presented is based on the assessment of conditional simulations that have incorporated the uncertainty in the interpretation of geological boundaries, tonnage, grade and consequently metal content estimates and likely production periods

The proposed classification technique does not replace the resource estimation; rather it serves as an additional tool to quantify confidence in the resource evaluation model.

It is further appreciated that particular Mineral Resource classification techniques are appropriate for specific situations.

A fundamental concept in Mineral Resource classification is the need for common sense and experience to prevail and it is therefore recommended that any automated mathematical technique applied be scrutinised.

9 References

Arik, A. (1999), An Alternative Approach to Ore Reserve Classification. *1999 APCOM Proceedings*, SME. Denver, p. 45-53

Blackwell, G (1998), Relative Kriging Errors – A Basis for Mineral Resource Classification, *Explor. Mining Geol.,* Vol.7 Nos 1 and 2, p. 99-105.

Dohm, C.E. (2003), Application of simulation techniques for combined risk assessment of both geological and grade models – an example. 2003 APCOM Proceedings, SAIMM, Cape Town, p. 351-354

Deutsch, C.V. and Journel, A.G., *GSLIB: Geostatistical Software Library and User's Guide*, 2nd Edition, Oxford University Press, 1998.

Goovaerts, P., *Geostatistics for Natural Resources Evaluation*, Oxford University Press, 1997.

Isaaks, E.H. and Srivastava, M.R., *An Introduction to Applied Geostatistics*, Oxford University Press, 1989.

Journel, A.G. and Huijbregts, C.J., *Mining Geostatistics*, Academic Press, 1978.

Snowden, D.V. (1996), Practical Interpretation of Resource Classification Guidelines, *AusIMM Annual Conference Procedings*, Perth March 24-28, p. 305-308.

Stephenson, P.R. and Stoker, P.T. (2001), Classification of Mineral Resources and Ore Reserves, *Mineral Resource and Ore Reserve Estimation, The AusIMM Guide to Good Practice,* Monograph 23, p. 653-659

Yamamoto, J.K. (2000), An Alternative Measure of the Reliability of Ordinary Kriging Estimates, *Mathematical Geology*, Vol32, No. 4, pp489-509.

This page intentionally blank

MINING

GEOSTATISTICS IN RESOURCE/RESERVE ESTIMATION: A SURVEY OF THE CANADIAN MINING INDUSTRY PRACTICE

MICHEL DAGBERT
Consultant, Geostat Systems International Inc.
10 Blvd de la Seigneurie E., Suite 203, Blainville, Qc, Canada, J7C 3V5

Abstract. With the new NI 43-101 rules of public disclosure for exploration and mining companies listed on the Canadian exchanges, it is now possible to have access to technical reports describing in details the procedures used by those companies to estimate resources and reserves for their properties in Canada and elsewhere in the world. This paper summarises the results of a survey of such technical reports issued in the last two years. It evaluates the role of geostatistics in various aspects of the resource/reserve estimation work namely the capping of outliers sample value, the domaining according to geology, the continuity analysis, the interpolation of block grades, the evaluation of dilution factors and the categorisation of resource/reserves.

1 Introduction

Since Feb. 1st, 2001, mining and exploration companies listed on Canadian stock exchanges (Toronto, Vancouver, Calgary and Montreal) must follow the so-called National Instrument (NI) 43-101 standard whenever they disclose technical and scientific information about their properties (CSA, 2000a and 2000b). Such information includes exploration results and of course resource and reserve estimates. Like similar standards in other countries(e.g., JORC,1999), NI43-101 does not specify how the actual exploration or estimation work should be done but concentrate on the profile of the individuals who do the work (the "qualified persons") and the format of the disclosure (the "technical reports"- CSA, 2000c). On the content itself, NI-43-101 endorses the revised resource/reserve definitions of the Canadian Institute of Mining and Metallurgy (CIM, 2000).

In the last 3.5 years of application of this new regulation, several hundreds "technical reports" of NI43-101 style have been filed. They are public documents available on the web in digital form and they constitute a privileged reference to determine how mineral resources and ore reserve are currently estimated in the Canadian mining industry and, in particular, to what extent geostatistics is used in this estimation

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 345-350.

2 Survey of technical reports

Technical reports are retrieved from the www.sedar.com site which concentrates all documents (annual reports, notices to shareholders, press releases...) in digital (PDF) form from public companies listed on Canadian stock exchanges. They are generally found in the "Other" category and are easily detected by their size of commonly several Mb since they correspond to documents of generally several hundred pages.

Exploration and mining companies are found in 3 industry groups: Gold and precious metals, Junior natural resource/Mining and Metals and Minerals. We have retrieved most of the technical reports issued since October 1, 2002 i.e. the last 2 years and all together they represent about 1200 documents.

About three quarter of those documents are uniquely concerned with so-called "exploration results" with limited drill hole information and no estimate of resource or reserve. That leaves about 300 reports dealing with properties with sufficient drilling/sampling data to warrant resource or reserve estimation (R&R reports).

The majority of the properties studied in those R&R reports is actually outside of Canada, mostly in United States, South America, Africa and Russia (with others in South Africa, China etc..). Gold is the most frequent commodity of interest (under the form of vein or disseminated type) followed by base metals (porphyry type and massive sulphide), uranium and industrial minerals. A surprisingly large fraction of properties in those R&R reports are producing mines (or have produced in the past), which can lead to some instructive reconciliation work.

Some of the issuers are well known Canadian mining houses like Barrick Gold Corp., Placer Dome Inc., Kinross Gold Corp. and Aur Resources Inc. Some of the properties described in those reports have made the headlines of mining publications in the recent years : Lac des Iles of North American Palladium Ltd., Kemess North of Northgate Minerals Corporation, Las Christinas of Crystallex International Corp.

As indicated in the introduction, the new NI 43-101 regulation puts a lot of emphasis on the "qualified person(s)" who authors the technical reports. They can be employees of the issuer but in most cases, they are outside consultants, independent of the issuer. Some of them are affiliated with large and well known consultancies from all over the world, particularly from Australia. It can be noted that there is a definite "correlation" between the approach taken in an R&R study and the background of the qualified person(s) responsible for that study and this is particularly true when it comes to the use of geostatistics.

The qualified person who authors a technical report is not necessarily the individual who has conducted the R&R work presented in the report. In many cases, the work has been done internally and the external consultant, who acts as the qualified person, after auditing and verification, simply endorses the results of that company work.

3 Use of geostatistics in resource and reserve estimation

In this age of widespread computer usage, a surprisingly high proportion (about 1/3) of the resource and reserve estimation work presented in the surveyed reports is still carried in the "manual" way with the interpretation of the limits of mineralised lenses from plans and sections and the calculation of an average grade of those lenses (or parts of them) from samples within those interpreted limits. This type of calculation seems restricted to vein type deposits to be mined by underground methods. Often there is a distinction between "geological" and "mining" resources, the later being after application of a minimum mining width to available vein intercepts and the corresponding grade dilution. Reserves are made of mining resources within limits of designed stopes and after application of dilution and recovery factors. Examples of such calculations can be found in Curtis, 2003 (for UG gold mines in Russia) and Roscoe et al, 2002 (for a UG gold mine in Canada). Needless to say, those R&R estimations do not use geostatistics.

The 2/3 balance of the R&R studies use the concept of computerized resource block model implemented in a mining package (Vulcan, Datamine, Gemcom, Mintec, Surpac…). Steps followed in such studies are invariably: geological solid modelling or "domaining", selection of block size, original sample compositing (with possibly some capping of high assays), eventually some variography of composite data in the various "domains", interpolation of block grade from surrounding composites (with search strategy and weighting scheme), categorization of estimated resources in blocks and finally conversion of resources to reserves.

Geological modelling consists of building 3D solids around material of the same "geological" nature based on lithology, alteration, degree of oxidation (the traditional leach/oxide/supergene/primary sequence of tropical or paleotropical terranes), geometry (blocks of similar general orientation in a folded structure) or simply grade. In the latter case, a low cut-off is used to delineate "potentially mineralised material" in disseminated mineralization, typically somewhere between 0.3 and 0.8 g/t in gold deposits. The resulting geological model can be fairly complex and detailed, for example 63 different solids for the Jinlonggou gold deposit in China (Fillis and Arnold, 2004). In the majority of cases surveyed, limits of those solids are "hard" limits i.e. blocks within limits are interpolated from just samples within the same limits. As a general rule, geostatistics is not used as an aid to geological modelling or to test the hard nature of the defined limits.

Resource model block size is quite variable and is linked to both the size of mineralised solids (small blocks in narrow zones) and the average spacing between samples (the old rule of the half distance between samples). It ranges from a high of 20x20x15m at Kemess North (Gray et al, 2004) to a low of 2x5x1m at the Barbrook mine (Applied Geology Service, 2004). Trend seems to have fairly small blocks especially with the sub-celling technique of mining packages (although generally all sub-cells in the same cell are given the interpolated value of that parent cell). In deposits to be mined by open-pit methods, the prevailing rule is to adjust the vertical dimension of blocks to the planned bench height of the future mine.

High grade capping of original assay data (whatever size of the corresponding interval) is the rule in gold deposits. The most popular approach to determine the cap limit is to look at changes in the slope of the cumulative frequency curve on log probability paper. In case of multiple domains, there is generally a specific cap for each domain which varies with the average grade of all samples in the domain. Unfortunately the proportion of capped samples as well as the percent gold metal lost is not always mentioned.

Composite size is generally dictated by the average size of the original assay intervals, irrespective of the dimension of blocks to be interpolated. As a result, composite size tends to be rather low, like 1m or 2m. In only a few instances, blocks for a deposit to be mined by open-pit are interpolated by bench composites (e.g. 5m composites for 10x10x5m blocks in 5m benches) thus minimizing risks of under-dilution of block grades.

Variography of composite grades in each domain is performed in only half the studies using resource block model. Description varies from an almost casual mention in the text to 120 pages of variogram plots (Belanger, 2003). Correlograms seem to prevail over regular variograms. We have not seen many indicator variograms or variograms of transformed data. The "pair wise relative variogram" is still very much popular with some consultants.

Inverse squared distance (ID2) is the most popular block grade interpolation method. We can even find cases where variograms are computed but blocks are interpolated by ID2 (Hill and Davidson, 2003). Arguments to prefer ID2 to kriging indicate that the later is not that well understood. For example : "*Inverse distance was used to interpolate gold grades into the block model for Bouroum instead of ordinary kriging due to the more significant presence of isolated high grade gold values that impacted adjacent low grade areas and the generally, more in-equidistant drill hole spacing at Bouroum*" (Vanin et al, 2003). When kriging is used, it is mostly under the form of ordinary kriging (OK) with very little indicator kriging (IK) applications.

Typical of this hesitation toward kriging is the report by Gosselin (2003) on Laronde gold deposit resources and reserves for 2003 : variograms are computed and fitted with anisotropic models in all 6 gold + base metals bearing sulphide zones but they are just used to defined search ellipsoids for the ID2 (or ID3) of blocks in the same zones.

Preference of ID2 over OK reflects some fear of diluting high-grade composites (or "smearing" high grades into adjacent low grade areas). This concern also transpires from the selection of search parameters and the fairly low maximum number of composites allowed in the interpolation (from 3 to 8 from the few studies where this information is given). Standard approach is to consider ellipsoids of increasing sizes to progressively fill the block matrix. The restricted search for high-grade composites option of some packages is sometimes used as an alternative (or in combination with) high grade capping prior to compositing.

In all the studies that we have surveyed, categorization of block model resources is strictly based on the geometry of composites with respect to blocks with no use of predicted uncertainty from variograms (the old kriging variances or the new conditional simulation). A common approach is to use the steps in the progressive search for composites around blocks to set the block category e.g. measured resources for blocks which can be interpolated with the most restrictive search (smallest ellipsoid, highest minimum number of composites..) up to inferred resources for the last blocks to be interpolated. In other cases, a specific template ellipsoid is used to test the density of composites around each block with cut-offs on number of composites within this ellipsoid corresponding to given drilling grid: for example, at Quebrada Blanca (Barr and Reyes, 2004), they use the number of 7.5m bench composites within a 75x75x18.75m ellipsoid to classify blocks with measured blocks if more than 20 composites (50x50m grid) and indicated blocks if more than 9 composites (100x50m grid).

Post-processing of block estimates is rather limited. In many cases, for deposits to be mined by open-pit methods, block values are used as-is in open-pit optimization and the calculation of reserves from resources (i.e. reserves are resources within final pit). In some instances, fixed recovery and dilution factors are applied to block estimates before pit optimization e.g. a 95%recovery in all blocks and a 10% dilution in contact blocks at Kemess North (Gray et all, 2004). In other cases, global change of support methods are used to check that block estimates have the grade variability corresponding to their size (Belanger, 2003). We have not found any study where conditional simulation is used to adjust block estimates to expected selectivity and grade control conditions of the future mining operation.

4 Conclusions

Roughly speaking, about one third of the resource and reserve studies issued by exploration and mining companies listed in Canada over the last two years use geostatistical methods in their resource estimation. Another third is also based on computerized block models but with inverse square distance. The last third is the manual approach with sectional blocks and sample averages within blocks.

Geostatistics used in those studies is fairly “classical” with composite grade variography and block estimation by ordinary kriging. Indicator kriging is not much used, even if the majority of studies deal with gold deposits. There is virtually no use of conditional simulation as an aid to resource categorization or block grade adjustment for recovery and dilution.

The general feeling that one gets when browsing those thousand pages of technical reports is that, at the moment, resource estimation is still more an art than a science with lots of subjective decisions and fudge factors (specially in high capping and categorization) which can be related to the background and past experience of the “qualified persons” who sign those reports.

Acknowledgements

Thanks are due to Gaetanne Beaulieu for her patience in retrieving technical reports on the web

References

Applied Geology Service, *Independent Qualified Person`s Report. Barbrook Mines Limited, Mpumalanga Province, South Africa* [Sedar : Caledonia Mining Corp.], 67p, 2004

Barr N.C. and Reyes R., *Report on Mineral Resources and Mineral Reserve Estimates at Dec. 31, 2003. Quebrada Blanca copper mine, Region I, Chile* [Sedar : Aur Resources Inc.], 57p, 2004

Belanger M., *Technical Report, La Coipa Mine, Chile* [Sedar : Kinross Gold Corp.], 231p, 2003

Cormier M., *Technical Mana Minéral S.A..Révision des Ressources Minérales, Mana , Burkina Faso, en date du 31 Decembre 2003* [Sedar : Semafo Inc.], 46p, 2004

CSA, *National Instrument 43-101. Standard of disclosure for mineral projects*, [available as *standards_disclosure_43-101-1.pdf* from www.osc.gov.on.ca], 21p, 2000a

CSA, *Companion policy 43-101CP to National Instrument 43-101. Standard of disclosure for mineral projects*, [available as *companion_policy_43-101.pdf* from www.osc.gov.on.ca], 14p, 2000b

CSA, *Form 43-101F1. Technical report*, [available as *technical_report_form_f1_43-101.pdf* from www.osc.gov.on.ca], 12p, 2000c

CIM, *CIM standards on mineral resources and reserves. Definitions and guidelines*, [available as *CIMdef1.pdf* from www.cim.org], 26p, 2000

Curtis L., *Technical Report. Zun-Holba and Irokinda Gold Mines, Buryatzoloto for High River Gold Mines Ltd.* [Sedar : High River Gold Mines Ltd], 17p, 2003

Fillis P. and Arnold C., *Tanjanshian Gold Project, Qinghai Province, 2003 work programme and resource estimation.* [Sedar : Afcan Mining Corporation], 54p, 2004

Gosselin G., *2003 Laronde Mineral Resource and Mineral Reserve Estimate Agnico-Eagle Mines Ltd., Laronde.* [Sedar : Agnico-Eagle Mines Ltd], 81p, 2003

Gray J.H., Morris R.J., Major K.W. and Arik, A., *Revised Kemess North pre-feasibility project* [Sedar : Northgate Exploration Ltd], 138p, 2004

JORC, *Australasian Code for Reporting of Mineral Resources and Ore Reserves*, [available as *JORC-code.pdf* from www.jorc.org],16p, 1999

Hill A. and Davidson A., *Technical Report. The Alto Chicama property. Department of La Libertad, Peruo.* [Sedar : Barrick Gold Corp.], 50p, 2003

Roscoe Postle Associates Inc., *Review of Mineral Resources and Mineral Reserves of the Macassa Mine Property, Kirkland Lake, Ontario, prepared for Kirkland Lake Gold Inc.* [Sedar : Kirkland Lake Gold Inc.], 73p, 2002

Salmon B., Mineral Resource and Mineral Reserve Estimates, Louvicourt Mine, January 1st, 2004, [Sedar : Aur Resources Inc.], 230p, 2004

Vanin D., Michaud M., Thalenhorst H., *Taparko-Bouroum project, Burkina-Faso. High River Gold Mines Ltd. 43 101 F1 technical report* [Sedar : High River Gold Mines Ltd], 153p, 2003

INTEGRATION OF CONVENTIONAL AND DOWNHOLE GEOPHYSICAL DATA IN METALLIFEROUS MINES

M. KAY, R. DIMITRAKOPOULOS and P. FULLAGAR
WH Bryan Mining Geology Research Centre
The University of Queensland, Brisbane Qld 4072, Australia

Geophysical logs provide valuable data that can be linked to orebody modelling and mine planning in metalliferous mines; however, geophysical measurements provide indirect data for ore grades and require further integration with conventional assay data. Integration can be based on the generation of suitable geophysical data compositing and the use of the sequential co-indicator simulation with the Markov-Bayes approximation. A detailed study at the Kidd Creek base metal mine, Canada, shows the practical aspects of the suggested approach and the value of integrating geophysical data.

1. Introduction

Geophysical logs provide valuable, relatively inexpensive information that can be further linked to various aspects of a mining operation, including orebody modelling, mine planning, grade control and production. Although financially attractive, downhole geophysical measurements usually only provide indirect indicators of ore grades (Fullagar and Fallon, 2001) and require further analysis in order to achieve integration with conventional assay data.

Until recently few studies have examined the issue of technical integration of geophysical data in the metalliferous environment. Early attempts include grade estimation based on natural gamma logs in uranium mines, after calibration of geophysically derived grades with geochemical assays (e.g. Bryan and Roghani, 1985; David, 1988). More recently, Miller and Luark (1993) use simulation techniques to construct models of rock strength in an underground coal mines; Dimitrakopoulos and Kaklis (2001) demonstrate the use of sequential co-indicator simulation to integrate data from a radio frequency electromagnetic tomography survey and scattered geochemical assays; and Basford et al. (2001) describe refinement of blasting based on automated interpretation of natural gamma and magnetic susceptibility logs. Kay (2001) provides a detailed study on integrating and valuing downhole geophysical measurements in base metal mines.

Given the potential economic benefits (e.g. Fallon et al., 1997) that could be obtained from using downhole geophysical logs, there is a clear incentive to determine how geophysical logging data can be integrated with orebody modelling and subsequently used throughout the mining process. To do this, it is necessary to use a simulation

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 351-361.

technique that is capable of integrating geophysical logs. This is a challenge since in the mine environment there is typically insufficient geophysical logging data to allow for calibration and variogram analysis. In addition, the handling of the variable support effects associated with geophysical signals in geologically complex environments has received little attention.

This paper demonstrates the effectiveness of technically integrating downhole geophysical logging in the metalliferous environment for orebody modelling. First, the sequential co-indicator simulation method with the Markov-Bayes approximation used herein is outlined. This is followed by a brief description of the approach employed to composite downhole geophysical logs. Subsequently, a case study with copper assays and conductivity log data from Kidd Creek Copper Mine, Canada, is used to illustrate the practical aspects of the technique.

2. Sequential co-indicator simulation in brief

Sequential co-indicator simulation (ScoIS) is a suitable method for integrating 'soft' data (such as geophysical logging measurements) with 'hard' data (such as conventional assay data) in simulating orebody models. ScoIS employs indicator co-kriging (Goovaerts, 1997) to derive the cumulative distribution function of the attribute being modelled and accounts for the geospatial correlation of 'hard' and 'soft' data, as well as their spatial cross-correlation. To alleviate the tedious inference and joint modelling of indicator covariance and cross-covariance functions, a variant of ScoIS employing the so-called Markov-Bayes approximation of Zhu and Journel (1992) is convenient, particularly when 'hard' and 'soft' data measure the same attribute. The Markov-Bayes hypothesis requires (i) that for a cut-off z_k a hard indicator datum (binary transform of the original measurement) *i(u, z_k)* at location **u** in a deposit prevails over the influence of any co-located soft indicator data *y(u, z_k)*; and (ii) that the indicator covariance function $C_{YY}(h;z)$ of the 'soft' data and cross covariance function $C_{YI}(h;z)$ between hard and soft data can be expressed as a function of the covariance of the hard data $C_{II}(h;z)$. Specifically,

$$\begin{aligned} C_{YY}(h;z) &= |B(z)| \cdot C_{II}(h;z) && \text{for } h = 0 \\ &= B^2(z) \cdot C_{II}(h;z) && \forall \text{h} > 0 \end{aligned} \tag{1}$$

and

$$C_{IY}(h;z) = B(z) \cdot C_{II}(h;z) \qquad \forall \text{h} \tag{2}$$

B(z) is an accuracy index given by:

$$B(z) = m^1(z) - m^0(z) \qquad B(z) \in [-1,1] \tag{3}$$

and

$$\begin{aligned} m^1(z) &= E\{Y(x;z) \mid I(x;z) = 1\} && m^1(z) \in [0,1] \\ m^0(z) &= E\{Y(x;z) \mid I(x;z) = 0\} && m^1(z) \in [0,1] \end{aligned} \tag{4}$$

where $m^1(z)$ and $m^0(z)$ are two conditional expectations that are obtained from calibration scatterplots of the hard versus the soft data, as shown in a subsequent section. It can be seen that *B(z)* is equal to one when the soft data are fully equivalent to the hard data.

3. Compositing downhole geophysical data

Geophysical logging data are usually collected every few centimetres downhole with the geophysical probe's sampling volume being quite variable. This variability of support is especially pronounced for conductivity measurements in base metal sulphide deposits due to the high conductivity of the ore. As a result, it is unclear how exactly geophysical logs should be composited. Kay (2001) uses an experimental approach which consists of compositing the conductivity logs with power averaging or l_r norm (e.g. Dimitrakopoulos and Desbarats, 1993; Claerbout and Muir, 1972). The power averaging exponent ω is derived experimentally. By varying the value of ω, one can generate a continuum of 'average' values that include common averages such as the arithmetic, geometric and harmonic means, when ω is equal to 1, 0 and –1 respectively. In determining a suitable ω value, it is rational to maximise the information content of the composited downhole geophysical measurements. The procedure for maximization has the following steps:

1. Composite, for a selected length, the geophysical log data at drillhole *X* with a particular power averaging value ω;
2. Estimate the metal grades in drillhole *X* using the composited conductivity log from Step 1 and the geochemical assays available from adjacent drillholes using standardised ordinary co-kriging (e.g. Deutsch and Journel, 1992);
3. Compare the resulting grade estimates to the actual geochemical composites in drillhole *X* using the mean squared error μ_{e^2}, the Spearman rank correlation coefficient *r'* (e.g. Swan and Sandilands, 1995) and the relative differences in these measures;
4. Repeat Steps 1 to 3 for all drillholes and different values of ω; and
5. Summarize all the results from Step 3 and choose a value for ω to use.

An example of this procedure is shown in a subsequent section.

4. Case Study: Integrating downhole conductivity logs and copper assays

Data collected from the Kidd Creek base metal deposit, Canada, illustrate the use of the above methods in integrating copper assay data and downhole inductive conductivity logs. Copper assays relate to 1.5m composites. Conductivity measurements were collected every five centimetres downhole and are composited to the same length of 1.5m using power averaging.

4.1. POWER AVERAGING OF GEOPHYSICAL LOGS

The procedure for deriving the power averaging constant ω described above is applied here to a set of drillholes. Kay (2001) showed that the accuracy of the copper estimates depends on the value of ω because μ_{e^2} has a minimum and *r'* a maximum in [-2.0, -0.5]. This suggests that ω should be drawn from [-2.0, -0.5]. But it is less clear which value of ω in [-2.0, -0.5] should be used since both the μ_{e^2} and *r'* maxima are quite broad in this interval (Kay 2001). However, if $\Delta\mu_{e^2}$ and $\Delta r'$ are examined instead (Figure 1), it is evident that ω be set at –1.0 (harmonic mean) since this value results in the minimum mean squared error and the maximum rank correlation coefficient. A series of such analyses for sets of drillholes with different dips suggested that an appropriate value of ω would be –1.0 (Kay, 2001). As a result all of the conductivity logs acquired in the study area were power average-composited using this value of ω and a sample interval of 1.5m. These composited logs are used for the simulation of the deposit.

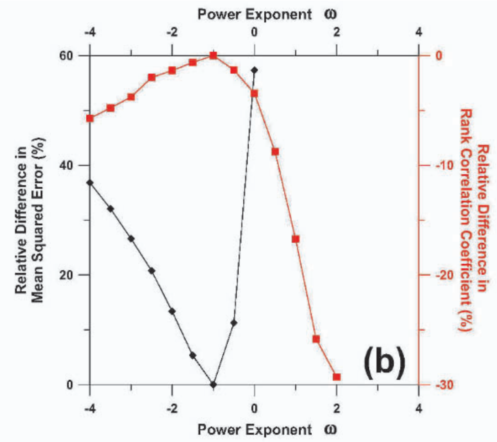


***Figure 1**. Relative difference in mean square error and rank correlation coefficient as a function of ω for composites generated from a cross-validation analysis of a set of representative drillholes (1.5m composite length).*

4.2. SIMULATION PARAMETERS

Having generated conductivity composites, copper grades are simulated with ScoIS and the Markov-Bayes approximation, which are conditional to the copper assays and conductivity logs. In this section, the practical aspects of the simulation method used and its assumptions are examined. ScoIS requires the selection of a set of cutoffs for the hard data (copper). The cutoffs were chosen to adequately characterise the overall copper data histogram as well as the metal content (Dimitrakopoulos, 2004). The latter is important in representing the copper metal quantity in the deposit, where 10% of the higher-grade samples may represent 50% of the metal in the deposit. Details of the application at Kidd Creek are included in Kay (2001).

The inference and modelling of indicator variograms for each copper cutoff involved two steps, namely identification of the three principal axes of the variogram ellipsoid, and modelling of experimental variograms along these three directions. For cutoffs above the 85th percentile, a different approach was followed (e.g., Dimitrakopoulos, 2004) where variogram parameters were experimentally adjusted to minimise order-relation problems. All indicator variograms were modelled using different linear combinations of the same set of basic structures (in this instance a nugget effect and an exponential model). Also, variogram parameters were varied smoothly from one cutoff to the next (Dimitrakopoulos, 2004; Goovaerts, 1997).

As previously described, the Markov-Bayes assumption allows the conductivity and copper-conductivity covariance models to be deduced from the copper covariance model and the calibration parameter $B(z_k)$ for each of the copper cutoffs z_k. However, this requires a set of cutoffs for the conductivity data to be specified. Eleven conductivity cutoffs were selected to ensure adequate characterisation of the declustered conductivity cumulative distribution function. This was achieved by selecting cutoffs that divided the conductivity values into classes of approximately equal frequency.

The resulting calibration parameters are presented in Figure 2. It can be seen that the calibration parameters are low for copper grades less than about 3.0%. However, for higher copper cutoffs, the calibration parameter stayed reasonably constant or increased slightly. These results are consistent with the general observations from the copper-conductivity scatter-plots, which suggest that the conductivity logs were a poor predictor of the copper grade in low-grade areas due to low signal level.

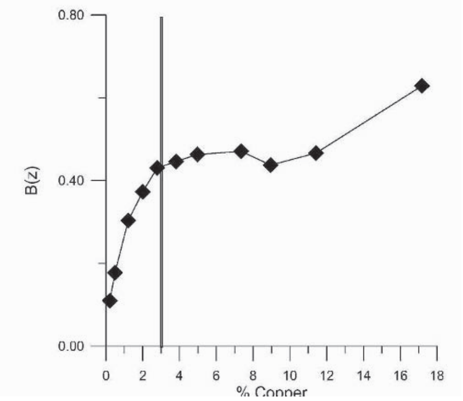


Figure 2. *Markov-Bayes calibration parameter versus copper grade for the conductivity logs.*

Also as previously discussed, conductivity and the copper-conductivity spatial models for each copper cutoff can be derived from the modelled copper-copper spatial model using a Markov-type hypothesis. Although there are no rigorous tests to verify the validity of this hypothesis, a useful check is to compare these derived variogram models with the corresponding experimental variograms. This check was performed on data from the test area for a range of copper cutoffs and is presented in Kay (2001). The results suggest that the derived models matched the spatial (auto) correlation associated with conductivity logs for low, median and high copper cutoffs.

In contrast, the cross correlation between the copper and conductivity was less well modelled using the Markov-Bayes assumption. Kay (2001) demonstrated that employing the Markov-Bayes assumption results in a variogram with a nugget to sill ratio of 0.80, whereas as the experimental data suggested that a value of 1.00 would be more appropriate. This effect was less marked for higher copper cutoffs, as the spatial correlation between copper and conductivity was quite well modelled for higher cutoffs (Kay, 2001). These results were further confirmation that conductivity logs were poor predictors of copper grade in areas with low copper concentrations. However, in spite of this shortcoming, it appeared that the Markov-Bayes hypothesis was valid.

The copper simulations generated in this study were produced using Zhu (1991)'s implementation of the Markov-Bayes simulation algorithm. In this paper the selection of the variogram parameters for cutoffs below the 20th and above the 85th percentile are explored. The reader is referred to Kay (2001) for an explanation of the other parameters. As discussed previously the variogram parameters for cutoffs between the 20th and 85th percentile can be inferred from manually fitting variograms. However, for cutoffs outside this interval, manually fitting variograms is unreliable since the associated experimental indicator variograms are very erratic. In the present study another approach was used to infer the variogram models for the extreme copper cutoffs. This approach consisted of the following steps:

1. Construct the experimental histogram (H^e) and indicator variograms (V^e) for all copper cutoffs using the declustered copper samples;
2. Infer variograms models (V^l) for the copper cutoffs above the 15th and below the 85th percentiles by manually fitting variogram models to V^e;
3. Assign starting values to the variogram models (V^h) for copper cutoffs below the 15th and above the 85th percentiles;
4. Generate copper simulations using the copper samples and the variogram models V^l and V^h;
5. Calculate the indicator variograms (V^{ml}) for cutoffs above the 15th and below the 85th percentile using the simulated copper values and compare these with the corresponding experimental indicator variograms V^e. If a visual inspection indicates that they are similar then proceed to Step 6. Otherwise adjust the manually fitted variograms V^l and return to Step 2;
6. Compare the histogram of simulated copper values (H^m) with that of the declustered copper assays for cutoffs that occur between the 15th and 85th percentiles. If they agree then proceed to Step 7, otherwise return to Step 2;
7. Compare the histogram of simulated copper values (H^m) with that of the declustered copper assays for the cutoffs below the 15th and those above the 85th percentiles. If they agree then proceed to Step 8, otherwise return to Step 3;
8. Examine the order relation corrections required during the generation of the simulations for cutoffs that occur between the 15th and 85th percentiles. If these are excessive, then return to Step 2, otherwise proceed to Step 9;
9. Examine the order relation correction required during the generation of the simulations for the cutoffs below the 15th and those above the 85th percentiles. If these are excessive, then return to Step 3, otherwise proceed to Step 10;
10. Stop - an acceptable set of copper indicator variograms has been generated.

The above procedure relies on determining how well the histogram of simulated copper values mimics the declustered histogram of copper assays. In the present study two measures were used. These were the *average difference* between the histogram of simulated values and the true experimental histogram, and the *average relative difference* between the two histograms. This latter quantity is important because, even though the high cutoffs only represent a very small proportion of the sample and simulated data, a small difference between the true and simulated histograms can be very economically significant. For example, an absolute difference of 1% between the two histograms may not be significant for the median class where 16% of the samples lie (i.e. 16±1%). However, the same absolute difference is highly significant for a high-grade class, where only 1% of the assay samples lie (i.e. 1±1%). The *average relative difference* reflects the relative importance of such differences. The procedure also requires that the number and magnitude of order relation corrections be examined in Steps 8 and 9 to determine whether an acceptable number of corrections were required, and this was selected to be an average magnitude of the probability corrections in the order of 0.01.

The iterative procedure described above was used to infer the variogram parameters for the high copper cutoffs in the study and resulted in a more finely tuned set of simulations (Kay 2001). This is shown in Figure 3 which presents the *average difference* and *average relative difference* for the set of simulations based on the final set of variogram parameters. The final set of variogram parameters resulted in realisations matching the experimental variograms very well for the high grade copper classes (Figure 4). With respect to the order relation corrections, the 'final' set of variogram parameters perform well (Kay, 2001).

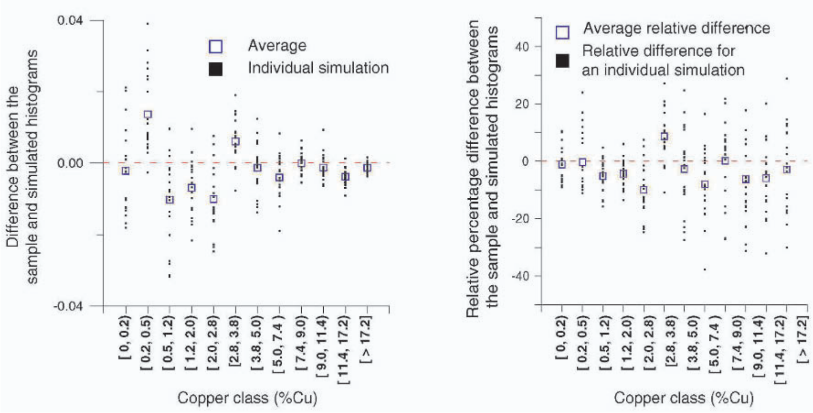


Figure 3. *Difference (left) and relative difference (right) between histograms of the copper composites and the ensemble of copper simulations. Also shown is the average copper simulation.*

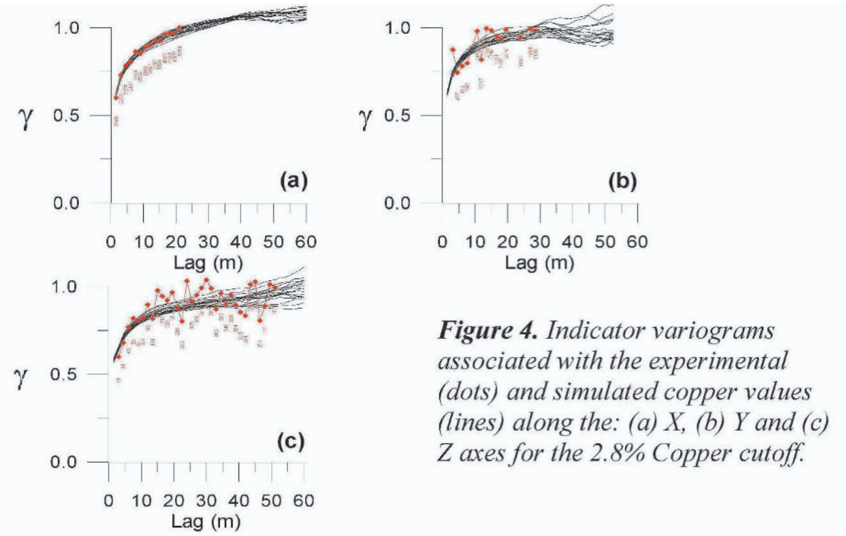


Figure 4. *Indicator variograms associated with the experimental (dots) and simulated copper values (lines) along the: (a) X, (b) Y and (c) Z axes for the 2.8% Copper cutoff.*

To show the effect of the contribution of conductivity data to the generation of copper simulations, three different groups of simulations are examined. Group A is the set of simulations generated using copper assays only. Group B was identical to A except that the copper assays associated with one drillhole fan are removed. The third set, Group C, has the same set of copper assays as Group B, but also includes the conductivity logs collected in the drillhole fan whose copper assay data had been excised from Group B.

Figure 5(a) presents a vertical section through a Group B simulation. The vertical section is in the plane where the excised drillhole fan was contained. With respect to the overall distribution of simulated copper grades, the Group A and B simulations appear to be similar (Kay, 2001). However, differences can be observed. Figure 5(b) displays some of the differences between a Group A and a Group B simulation that both use the same random seed. By removing the copper assays in the drill fan the resulting Group B simulation has considerably higher grades than the Group A simulation in the mineralised zone centred at X=275 (ie the zone outlined in red). However, as Figure 5(b) indicates, the opposite is true in the second mineralised zone (i.e. the zone outlined in black). The figure suggests that removing the copper assays in this zone results in much lower simulated copper grades.

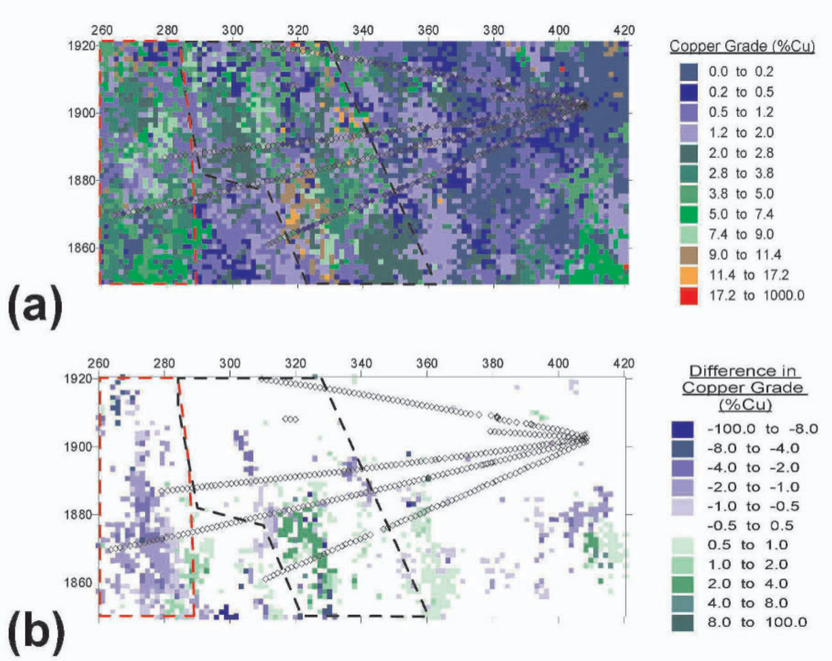


Figure 5. *Vertical section through a representative Group B simulation (a) simulated Group B copper values; and (b) difference between the Group C and corresponding Group A simulated copper values. Also shown are the sample locations associated with the excised drillhole fan.*

Figure 6 illustrates the effect of using conductivity logs in the simulation process. It is evident that the Group C simulations, using the conductivity logs collected in the drillfan, are qualitatively similar to the A and B simulations. For example, the Group C simulation contains the two mineralised zones discussed earlier. However, Figure 6(b) shows that using the conductivity logs results in simulations that are very similar to the Group A simulations. For instance in the plane of the vertical section, the difference between the equivalent Group A and C simulations is less than ±0.5% Cu for more than 90% of the vertical section. Moreover, in the remaining 10% of the section, the difference is less than ±1.0% Cu with only a few simulated points differing by more than ±5.0% Cu. This suggests that, in the plane of the vertical section using conductivity logs results in simulations that are very similar to those based purely on copper assays.

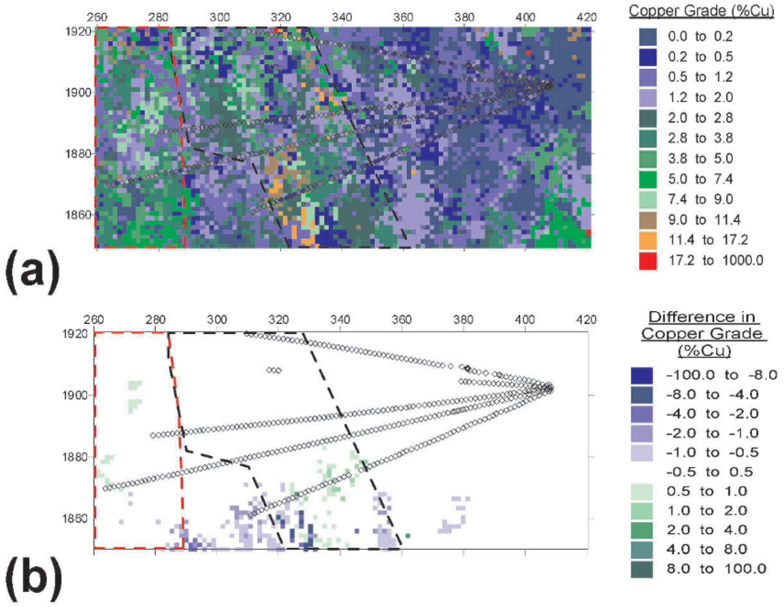


Figure 6. *Vertical section through a representative Group C simulation (a) simulated Group C copper values; and (b) difference between the Group C and corresponding Group A simulated copper values. Also shown are the sample locations associated with the excised drillhole fan.*

5. Conclusions

The ability to technically integrate geophysical data with conventional assays in metalliferous mines is important during the developmental and mining stages where orebody models are constructed. The present study described the conditional simulation of ore grades in an application that integrates drill core assay data and downhole geophysical logs. This application was used to assess the variability of copper grades at the Kidd Creek base metal mine, Canada. Composites were generated using generalised power averaging which aims to maximise extraction of information from the conductivity logs. Sequential co-indicator simulation was applied to 'hard' (copper assay) and 'soft' (conductivity logging) data using the Markov-Bayes approximation. The selection of the cutoff grade was based on the quantity of metal, and the iterative calibration of indicator variograms at very high cutoffs was performed to ensure convergence with declustered copper assay statistics. Validation of the simulations suggests the Markov-Bayes approximation works reasonably well. It was shown that simulated realisations of copper grades are of comparable quality, when some of the assay composites in selected drillholes are replaced by conductivity data. This suggests that replacing some assaying with logging can generate savings with little loss of information.

Acknowledgments

The support of the Kidd Creek Mine, Ontario, Canada and Falconbridge Ltd is gratefully acknowledged.

References

Basford, P., Kelso, I., Briggs, T., Clifford, M., Anderson, R., and Fullagar, P., 2001, Development of a Short-term Model using Petrophysical Logging at Century Mine, North Queensland: *Australian Society of Exploration Geophysicists Preview*, No. 92, 19-24.

Bryan, R.C. and Roghani, F., Application of Conventional and Advanced Methods to Uranium Ore Reserve Estimation and the Development of a Method to Adjust for Disequilibrium Problems. *17th APCOM Symposium,* 1985, p. 109-120.

Claerbout, J. and Muir. F., Robust Modeling With Erratic Data. *Geophysics*, vol. 38, 1973, p. 826-844.

David, M., *Handbook of Applied Advanced Geostatistical Ore Reserve Estimation.* Elseveir, New York, 1988.

Deutsch, C. and Journel, A., *GSLIB Geostatistical Software Library and Users Guide*. Oxford University Press, New York, 1992.

Dimitrakopoulos, R., *Risk Assessment In Orebody Modelling And Mine Planning: Decision-making with uncertainty*. BRC Notes, SME Annual Meeting & Exhibit, Denver Colorado, 2004, p. 300.

Dimitrakopoulos, R. and Desbarats, A., Geostatistical Modeling of Gridblock Permiabilities for 3D Reservoir Simulators, *SPE Reservoir Engineering,* 1993, p. 13-18.

Dimitrakopoulos, R. and Kaklis, K., Integration of Assay and Cross-hole Tomographic Data in Orebody Modelling: Joint Geostatistical Simulation and Application at Mount Isa Mine, *Queensland. Transactions, IMM*, 110 (Jan.-April), 2001, p. B33-B39.

Fallon, G., Fullagar, P. and Sheard, S., Application of Geophysics in Metalliferous Mines, *Australian Journal of Earth Sciences*, vol. 44, no. 4, 1997, p. 391-409.

Fullagar, P.K., and Fallon, G.N., 2001, Geophysical GradeEstimation at Mines, *Australian Society of Exploration Geophysicists Preview*, No. 90, 30-32.

Goovaerts, P., *Geostatistics for Natural Resources Evaluation.* Oxford University Press, New York, 1997.

Kay, M., Geostatistical Integration of Conventional and Downhole Geophysical Data in the Metalliferous Mine Environment*.* MSc thesis, University of Queensland, 2001.

Miller, S. and Luark, R., Spatial Simulation of Rock Strength Properties Using a Markov-Bayes Method, *International Journal of Rock Mechanics and Mining Science and Geomechanics Abstracts,* vol. 30, no 7*,* 1993, p. 1631-1637.

Swan, A. and Sandilands, M., *Introduction to Geological Data Analysis*. Blackwell Science, Oxford, 1985.

Zhu, H., *Modeling Mixture of Spatial Distributions with Integration of Soft Data.* Phd Thesis, Stanford University, 1991.

Zhu, H. and Journel, A., Formatting and Integrating Soft Data: Stochastic Imaging via the Markov-Bayes Algorithm: in Soares, A. (Ed.) *Geostatistics Troia '92*. Kluwer Academic Publishers, Dordrecht, 1992, p. 1-12.

This page intentionally blank

THE KRIGING OXYMORON: A CONDITIONALLY UNBIASED AND ACCURATE PREDICTOR (2nd EDITION)

EDWARD ISAAKS, Ph.D.
ISAAKS & Co.

Abstract. An analysis of conditional bias and its impact on mineral resource estimation is presented. A simple method is proposed for building a long-term mineral resource block model that accounts for conditional bias, change-of-support, and the information effect at the time of mining.

1. Introduction

Accounting for change-of-support, the information effect, and conditional bias are problems well known to mineral resource modelers. Although the methods proposed for dealing with change-of-support and the information effect are little more than approximations, case studies suggest these methods are useful (David, 1977; Journel and Huijbregts, 1978; Matheron, 1984; Parker, 1980; Isaaks and Srivastava, 1991; Deraisme, 2000). However, the same cannot be said for conditional bias. A literature review reveals that conditional bias is poorly understood and that many of the claims are misleading.

Krige (1994; 1996; 1999) claims the preliminary prerequisite of all resource estimators is the elimination of conditional bias. Sinclair and Blackwell (2002) claim that conditional bias contributes to the discrepancies noted between the prediction of recoverable resources and production. David, Marcotte, and Soulie (1984) propose a correction for conditional bias and claim that this correction will reduce the discrepancy between predicted resources and production. Pan (1998) proposes a correction for conditional bias followed by a correction for the smoothing induced by the first correction. However, it can be shown that these two corrections are circular in the sense that the final smoothing correction re-introduces conditional bias. Guertin (1984) proposes a solution that she claims can be easily implemented as a correction factor for *any mineral resource estimation* or grade control system. Deutsch and McLennan (2003) argue that conditionally simulated block model values are both conditionally unbiased and accurate predictors of the tons and grade that will be recovered at the time of mining. However, as will be shown these claims are not correct despite their wide acceptance.

A conditionally unbiased and accurate predictor[1] is an oxymoron. The estimator for a long-term mine planning block model may be conditionally unbiased but then the

[1] Accuracy is defined as the ability of the long-term block model to predict the actual tonnage and average ore grade that will be recovered at the time of mining.

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 363-374.

histogram of block estimates will be smoothed yielding inaccurate predictions of the recoverable tons and grade above cutoff grade. Conversely, if the histogram of block estimates provides accurate predictions, then the block estimator is necessarily conditionally biased. The estimator for a long-term mine planning block model cannot be conditionally unbiased and simultaneously accurate as claimed by Deutsch and McLennan (2003). David (1977) recognized the oxymoron by pointing out that one can accurately estimate the histogram of block grades but then one cannot localize the blocks. Alternatively, one can estimate as accurately as possible the grades of precisely located blocks (thereby minimizing conditional bias) but then the block histogram will be smoothed. The only exception to this apparent contradiction occurs when the block estimates are perfectly correlated with the true block grades. In this unlikely scenario the block model is both a conditionally unbiased and accurate predictor.

This paper provides an analysis of conditional bias and its impact on mineral resource estimation. Although it may not be possible to eliminate conditional bias from the grade control estimator it can be evaluated through conditional simulation. Block models for long-term mine planning can be built using simulation methods that not only quantitatively account for the conditional bias of future grade control estimators, but also for future change-of-support and information effects.

Section 2 defines two types of mineral resource block models on the basis of how the block estimates are used by the mine. These definitions provide the key to understanding the role of conditional bias in mineral resource modeling. Section 3 provides a formal definition of conditional bias and describes a simple check. Section 4 examines the impact of conditional bias on prediction when the block estimates are used for selection at the time of mining e.g., grade control. Section 5 examines the problem of predicting the tons and grade that will be recovered at the time of mining given that selection will be made using grade control estimates based on future blast hole data. Section 6 describes how to build a long-term mine planning block model by conditional simulation that accounts for a future conditionally biased grade control estimator, the information effect, and a change of support.

2. Two Types of Block Models

Mineral resource models can be classified into one of two types depending on how the block estimates are used by the mine operation.

Type 1: Models whose block grade estimates are used to predict the tons and average grade of ore material that will be recovered each annual, semi-annual, or quarterly period over the life-of-mine are classified as Type 1. Individual block estimates are typically derived from relatively sparse diamond drill hole (DDH) data. Predicted recoveries made from Type 1 estimates are useful for feasibility studies, long and short term mine planning, and the estimation of production schedules etc. Individual block estimates are not used for selection at the time of mining. Thus, it is not necessary to know the precise location or recoverable grade of each ore block. Knowledge of the distribution of recoverable[2] block grades to be mined in the future for each period is sufficient. Type 1 models are often referred to as long-term (mine planning) models.

Type 2: Models whose block grade estimates are used for selection at the time of mining are classified as Type 2. Individual block volumes are equivalent to a selective mining unit (SMU) with the grade of each block typically estimated from neighboring blast hole (BH) grades. The use of these estimates to distinguish between ore and waste is commonly known as *grade control*.

3. Definition of Conditional Bias

3.1 NOTATION

D : The deposit or domain of interest.
$[Z(\mathbf{u}),\ \mathbf{u} \in D]$: A stationary random function consisting of a set of point support random variables.

$Z_v(\mathbf{u}) = \frac{1}{|v|} \int_{v(\mathbf{u})} Z(\mathbf{u}')d\mathbf{u}'$: A random variable of support v centered at location $\mathbf{u}$.

$[Z_v(\mathbf{u}), \mathbf{u} \in D]$: A stationary random function consisting of a set of random variables of support *v*. The random function $[Z_v(\mathbf{u}), \mathbf{u} \in D]$ is written as Z_v to simplify notation.
$F_v(z;\mathbf{u} \mid (n)) = \text{prob}\{Z_v(\mathbf{u}) \le z \mid (n)\}$: Non stationary cumulative conditional distribution function (ccdf) of the random variable $Z_v(\mathbf{u})$ at the location $\mathbf{u}$ conditioned by n data.

$F_D(z;v \mid (n)) = \frac{1}{|D|} \int_D F_v(z;\mathbf{u}' \mid (n))d\mathbf{u}'$: The probability that the grade of a randomly selected SMU within the domain D will be no greater than the cutoff *z*.
$[Z_{v*}(\mathbf{u} \mid (n)), \mathbf{u} \in D]$: A non stationary random function consisting of a set of random variables where each RV $Z_{v*}(\mathbf{u} \mid (n))$ is of the form $\sum_{i=1}^{n} w_i\ Z(\mathbf{u}_i)$ with $\sum w = 1$. The random function $Z_{v*}(\mathbf{u} \mid (n)),\ \ \mathbf{u} \in D$ is written as Z_{v*} to simplify notation.
$F_{v*}(z;\mathbf{u} \mid (n)) = \text{prob}\{Z_{v*}(\mathbf{u} \mid (n)) \le z\}$: The non-stationary ccdf of the random variable $Z_{v*}(\mathbf{u})$ at the location $\mathbf{u}$ conditioned by the *(n)* data.

$F_D(z;v^* \mid (n)) = \frac{1}{|D|} \int_D F_{v^*}(z;\mathbf{u}' \mid (n))d\mathbf{u}'$: The probability that the estimated grade $Z_{v*}(\mathbf{u})$ of a SMU randomly selected within *D* will be no greater than *z*.

3.2 DEFINITION

The conditional expectation is given by:

$$E\{Z_v \mid Z_{v*} = z\} = h(z) \qquad \forall z \tag{1}$$

[2] The recoverable grade is the actual grade recovered given that selection is based on estimates typically made from blast hole data at the time of mining.

where the function $h(z)$ may be linear or non linear. However, if we impose the condition:

$$h(z) = z \qquad \forall z \tag{2}$$

the function $h(z)$ will be linear through the origin with a slope of 1.0 and the estimator Z_{v*} is conditionally unbiased by definition (Journel and Huijbregts, 1978).

The conditionally unbiased relation (2) can be re-written as:

$$E\{Z_v - Z_{v*} \mid Z_{v*} = z\} = 0 \qquad \forall z \tag{3}$$

Equations (2) and (3) imply that the average of the estimator Z_{v*} above cutoff is an unbiased estimate of the average of the corresponding true values Z_v:

$$E\{Z_v \mid Z_{v*} > z\} = E\{Z_{v*} \mid Z_{v*} > z\} \qquad \forall z \tag{4}$$

Equation (4) can also be written as:

$$E\{Z_v - Z_{v*} \mid Z_{v*} > z\} = 0 \qquad \forall z \tag{5}$$

3.3 A CHECK FOR CONDITIONAL BIAS

The linear regression of Z_v on Z_{v*} is given by $E\{Z_v \mid Z_{v*} = z\} = a*z + b$ where *a* is the slope and *b* the intercept. Thus, if the slope of the linear regression of Z_v on Z_{v*} is not equal to 1 or the intercept is not equal to 0, then the estimator Z_{v*} is conditionally biased, e.g.,

$$\begin{aligned} E\{Z_v \mid Z_{v*} = z\} &= h(z) \\ a*z+b &= z,\ \forall z \quad \text{iff } a=1 \text{ and } b=0 \end{aligned} \tag{6}$$

The linear regression model also provides some insight on the relationship between the two random functions Z_v and Z_{v*} e.g., the slope *a* is given by:

$$a = \frac{\operatorname{cov}(Z_v Z_{v*})}{\operatorname{var}(Z_{v*})} = \frac{\sigma_v}{\sigma_{v*}} \rho_{v\,v*} \tag{7}$$

where σ_v^2 and σ_{v*}^2 are the variances of Z_v and Z_{v*}. Thus, for a conditionally unbiased estimator:

$$a = \frac{\sigma_v}{\sigma_{v*}} \rho_{vv*} = 1.0 \tag{8}$$

Two important observations can be made from Equation (8).

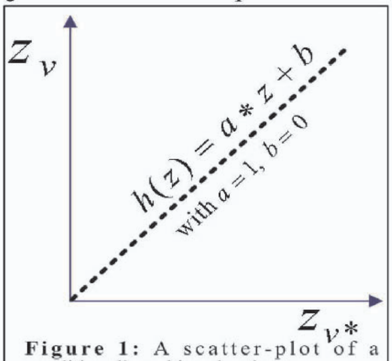


Figure 1: A scatter-plot of a conditionally unbiased estimator.

1. Since in practice, the correlation between the true and estimated values is: $\rho_{v\,v*} < 1$, then for a conditionally unbiased estimator, necessarily: $\sigma_v^2 > \sigma_{v*}^2$. In other words, the estimates of a conditionally unbiased estimator are smoothed

.

2. Conversely, if the two distributions $F_D(z; v \mid (n))$ and $F_D(z; v^* \mid (n))$ defined in section 3.1 have equal variances: $\sigma_v^2 = \sigma_{v^*}^2$, then necessarily $a < 1$. That is, the estimator Z_{v^*} is conditionally biased.

4. Type 2 Estimates and their Recovery Functions

Recall that the estimates of a Type 2 estimator are used by the mine operation for the selection of ore at the time of mining.

4.1 NOTATION

The type 2 estimator is denoted by a double asterisk, e.g., $Z_{v^{**}}(\mathbf{u})$ **.**

$F_{v^{**}}(z; \mathbf{u} \mid (n))$: The non stationary ccdf of the RV $Z_{v^{**}}(\mathbf{u})$ at location $\mathbf{u}$ conditioned by the (n) data.

$F_D(z; v^{**} \mid (n)) = \frac{1}{|D|} \int_D F_{v^{**}}(z; \mathbf{u}' \mid (n)) d\mathbf{u}'$: The non stationary conditional probability that the estimated grade $Z_{v^{**}}(\mathbf{u})$ of a randomly selected SMU within the domain D will be no greater than z. Note that this distribution is commonly estimated in practise.

$F_v(z; \mathbf{u} \mid v^{**}, (n))$: The non stationary ccdf of the RV $Z_v(\mathbf{u})$ at location $\mathbf{u}$ given that $Z_{v^{**}}(\mathbf{u}) \leq z$ and the (n) conditioning data.

$F_D(z; v \mid v^{**}, (n)) = \frac{1}{|D|} \int_D F_v(z; \mathbf{u}' \mid v^{**}, (n)) d\mathbf{u}'$: The non stationary conditional probability that the true grade $Z_v(\mathbf{u})$ of a randomly selected SMU within the domain D will be no greater than z given that its estimated grade $Z_{v^{**}}(\mathbf{u})$ is no greater than z. Note that this distribution is not known nor is it commonly estimated in practice.

4.2 ACTUAL RECOVERIES GIVEN THAT SELECTION IS MADE USING ESTIMATED GRADES.

The following recovery functions describe the *actual but unknown* quantities that will be recovered given that selection is made using the estimates $z_v^{**}(\mathbf{u})$ **.** The recovered tonnage is given by:

$$T_D(z) = T_o[1 - F_D(z; v^{**} \mid (n))] \qquad \forall z \tag{9}$$

The actual but unknown quantity of recovered metal is given by:

$$Q_D(z) = T_o \int_z^{\infty} z' \, dF_D(z'; v \mid v^{**}, (n)) \qquad \forall z \tag{10}$$

The actual but unknown recovered grade is given by:

$$m_D(z) = \frac{Q_D(z)}{T_D(z)} \tag{11}$$

4.3 ESTIMATED RECOVERIES GIVEN THAT SELECTION IS MADE USING ESTIMATED GRADES

The recovery equations provided by (10) and (11) are not useful since the distribution $F_D(z; v \mid v^{**}, (n))$ is not known or commonly estimated in practice. However, by replacing the unknown distribution with the commonly estimated distribution $F_D(z; v^{**} \mid (n))$, one can estimate the recoveries as follows:

$$\hat{T}_D(z) = T_o[1 - F_D(z; v^{**} \mid (n))] \qquad \forall z \tag{12}$$

The estimated recovered quantity of metal is given by:

$$\hat{Q}_D(z) = T_o \int_z^\infty z' \, dF_D(z'; v^{**} \mid (n)) \qquad \forall z \tag{13}$$

and the estimated recovered grade is given by:

$$\hat{m}_D(z) = \frac{\hat{Q}_D(z)}{\hat{T}_D(z)} \tag{14}$$

If the estimator $Z_{v^{**}}(\mathbf{u})$ is conditionally unbiased, then the estimated recoveries (13) (14) will be equal to the actual recoveries (10) (11) since conditional unbias implies the following:

$$\begin{aligned}
& E\{Z_v \mid Z_{v^{**}} > z\} && = E\{Z_{v^{**}} \mid Z_{v^{**}} > z\} \\
\Rightarrow \quad & \frac{\int_z^\infty z' dF_D(z'; v \mid v^{**}; (n))}{1 - F_D(z; v^{**} \mid (n))} && = \frac{\int_z^\infty z' dF_D(z'; v^{**} \mid (n))}{1 - F_D(z; v^{**} \mid (n))} \qquad \forall z \qquad (15) \\
\Rightarrow \quad & \int_z^\infty z' dF_D(z'; v \mid v^{**}; (n)) && = \int_z^\infty z' dF_D(z'; v^{**} \mid (n)) \\
\Rightarrow \quad & F_D(z'; v \mid v^{**}; (n)) && = F_D(z'; v^{**} \mid (n))
\end{aligned}$$

Thus, it appears[3] that the estimator $Z_{v^{**}}(\mathbf{u})$ must be conditionally unbiased in order to provide accurate predictions of the tons and grade that will be delivered to the mill. Ideally, the estimator $Z_{v^{**}}(\mathbf{u})$ will also minimize the conditional variance $E\{[Z_v - h(z)]^2\}$ (Journel and Huijbregts, 1978) so as to minimize ore loss and dilution or misclassification at the time of mining.

5. Type 1 Estimates and their Recovery Functions

Recall, that Type 1 estimates are used to predict the tons and grade of ore that will be recovered in the future at the time of mining. They are not used for selection at the time of mining.

[3] Section 6 describes how conditional simulation can be used to accurately predict the tons and grade that will be delivered to the mill in spite of a conditionally biased grade control estimator.

5.1 NOTATION

Type 1 estimates are denoted by a single asterisk, e.g., $z_{v*}(\mathbf{u})$.

$F_{v*}(z;\mathbf{u}\,|\,(n))$: The non stationary ccdf of the RV $Z_{v*}(\mathbf{u})$ at location $\mathbf{u}$ conditioned by the (n) data.

$F_D(z;v^*\,|\,(n)) = \frac{1}{|D|}\int_D F_{v*}(z;\mathbf{u}'\,|\,(n))d\mathbf{u}'$: The non stationary conditional probability that the estimated grade $Z_{v*}(\mathbf{u})$ of a randomly selected SMU within the domain D will be no greater than z. Note that this distribution is commonly estimated in practice.

5.2 THE RECOVERY EQUATIONS

Recoverable tonnage:

$$\widehat{T}_D(z) = T_o[1 - F_D(z;v*\,|\,(n))] \qquad \forall \mathrm{z} \tag{16}$$

Recoverable quantity of metal:

$$\widehat{Q}_D(z) = T_o \int_z^{\infty} \mathrm{z}'\, dF_D(z';v*\,|\,(n)) \qquad \forall z \tag{17}$$

Recoverable grade:

$$\widehat{m}_D(z) = \frac{\widehat{Q}_D(z)}{\widehat{T}_D(z)} \tag{18}$$

Recall, that (9), (10), and (11) provide the actual recoveries given that selection will be made using the estimates Z_{v**} in the future. Thus, to be useful the recoveries predicted by (16), (17), and (18) must be equal to those given by (9), (10), and (11). However, this is a problem since there is nothing in (16), (17), and (18) that guarantees equivalence to (9), (10), and (11). This problem is recognized within the mining industry where a common solution is to impose additional constraints on the estimators $z_{v*}(\mathbf{u})$ and $z_{v**}(\mathbf{u})$, e.g.,

Condition 1. $F_D(z;v*\,|\,(n)) = F_D(z;v**\,|\,(n'))$ - This condition requires the histogram of the type 1 estimates to be equal to the histogram of the type 2 estimates within *D*. For example;

- In practice, the future distribution $F_D(z;v**\,|\,(n))$ is estimated using smoothing relations and the change of support hypothesis (Journel and Huijbregts, 1978; Isaaks and Srivastava, 1989; Sinclair and Blackwell, 2002).
- The distribution $F_D(z;v*\,|\,(n))$ is then made to match as close as possible to the estimated distribution $F_D(z;v**\,|\,(n))$ by controlling the number of samples used to estimate z_{v*} locally (Deutsch and McLennon, 2003).

Condition 2. $E\{Z_v - Z_v^{**}\,|\,Z_v^{**} > z\} = 0 \ \ \forall z$ - This condition requires the type 2 estimator to be conditionally unbiased.

The equivalence between the predicted recoveries (16), (17), and (18) given conditions (1) and (2) and the actual recoveries (9), (10), and (11) is easily confirmed.

However, condition (1) may not be that easy to impose on the estimator Z_{v*}. The change of support and information effect may render the distributions $F_D(z;v^*|(n))$ and $F_D(z;v^{**}|(n))$ incomparable. Thus, at best this practice amounts to nothing more than an approximation.

Condition (2) may also be difficult if not impossible to impose on the future estimator Z_{v**}. Although kriging is said to be a *conditionally unbiased estimator,* in reality it is conditionally unbiased if and only if the distribution of $Z(\mathbf{u})$ is normal and its mean $E\{Z(\mathbf{u})\}$ is known (David, 1977). The problem is that almost all distributions of $Z(\mathbf{u})$ in mining applications are non-normal with relatively large coefficients of variation and large coefficients of skew. Because of this, it is very difficult if not impossible for the mine operator to insure that the grade control estimator is conditionally unbiased.

5.3 THE OXYMORON

Note, that although the estimator Z_{v*} is an accurate predictor of recoveries (9), (10), and (11) given conditions (1) and (2), Z_{v*} is almost certain to be conditionally biased. For example, from condition (2),

$$\frac{\sigma_v}{\sigma_{v**}}\rho_{vv**} = 1.0 \tag{19}$$

and from condition (1),

$$\sigma_{v*} = \sigma_{v**} \tag{20}$$

and since $\rho_{vv*} < \rho_{vv**}$ with near certainty then,

$$a = \frac{\sigma_v}{\sigma_{v*}}\rho_{vv*} < 1.0 \tag{21}$$

that is, Z_{v*} is almost certain to be conditionally biased. Thus, in spite of conditional bias, Z_{v*} may be an accurate predictor of recoverable resources given conditions (1) and (2).

6 Conditional Simulation and Prediction

This section proposes a method for building the long-term block model using conditional simulation via the LU decomposition of the covariance matrix, (Davis, 1987).

6.1 NOTATION

$Z_{\tilde{v}}(\mathbf{u})$ - the tilde above a variable denotes a conditionally simulated value. Otherwise the notation for the simulated variables and their distributions is identical to the definitions provided in section 4.1

6.2 CONDITIONAL SIMULATION

Consider the following vectors of point support Gaussian random variables:

$\mathbf{Y}_1 = [Y(\mathbf{u}_i), i = 1, n]'$ - a vector of (n) $N(0,1)$ random variables located at DDH sample locations $\mathbf{u}_i \;\; i = 1, n$,

$\mathbf{Y}_2 = [Y(\mathbf{u}_j), j = 1, s]'$ - a vector of (s) $N(0,1)$ random variables located at blast hole (BH) locations $\mathbf{u}_j \;\; j = 1, s$, and

$\mathbf{Y}_3 = [Y(\mathbf{u}_k), k = 1, t]'$ - a vector of (t) $N(0,1)$ random variables located at the discretization point locations $\mathbf{u}_k \;\; k = 1, t$ of the SMU.

Note that some of the locations may be co-located, e.g., $\mathbf{u}_i = \mathbf{u}_j$, $\mathbf{u}_i = \mathbf{u}_k$, $\mathbf{u}_j = \mathbf{u}_k$ for some i, j, k (see Figure 2).

The corresponding covariance matrices are given by:

$\mathbf{C}_{11} = \text{cov}(\mathbf{Y}_1\mathbf{Y}_1')$ with dimension $n \; x \; n$

$\mathbf{C}_{21} = \text{cov}([\mathbf{Y}_2', \mathbf{Y}_3']'\mathbf{Y}_1')$ with dimension $m \; x \; n$ where $s + t = m$.

$\mathbf{C}_{22} = \text{cov}([\mathbf{Y}_2', \mathbf{Y}_3']'[\mathbf{Y}_2', \mathbf{Y}_3'])$ with dimension $m \; x \; m$.

The covariance matrix between the random vectors $\mathbf{Y}_1, \mathbf{Y}_2$, and $\mathbf{Y}_3$ can be decomposed into the product of a lower and upper triangular matrix, e.g.,

$$\left[\begin{array}{c|c} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \hline \mathbf{C}_{21} & \mathbf{C}_{22} \end{array}\right] = \left[\begin{array}{c|c} \mathbf{L}_{11} & \mathbf{0} \\ \hline \mathbf{L}_{21} & \mathbf{L}_{22} \end{array}\right] * \left[\begin{array}{c|c} \mathbf{U}_{11} & \mathbf{U}_{21} \\ \hline \mathbf{0} & \mathbf{U}_{22} \end{array}\right] \tag{22}$$

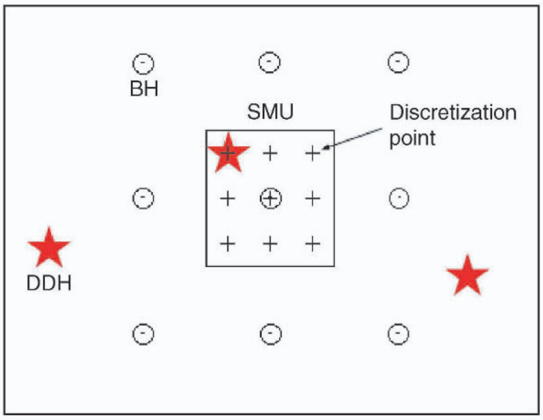


Figure2: Example locations of the random variables $\mathbf{Y}$ relative to a SMU. The stars represent $\mathbf{Y}_1$ at DDH locations, while the circles represent $\mathbf{Y}_2$ at BH locations and the plus signs symbolize $\mathbf{Y}_3$ at the discretization points of the SMU. Note the co-location of some of the variable locations.

Next, we interpret the relatively sparse DDH data $z(\mathbf{u}_i),\ i=1,n$ as a realization of the random vector $\mathbf{Y}_1$, e.g.,

$$y_1(\mathbf{u}_i)=\varphi(z(\mathbf{u}_i)),\ i=1,n \tag{23}$$

where $\varphi(\cdot)$ is the normal score transform. Realizations of the random vectors $\widetilde{\mathbf{Y}}_2$ (at BH locations) and $\widetilde{\mathbf{Y}}_3$ (at SMU discretization point locations) can be simulated conditional to the transformed DDH data $\mathbf{Y}_1$ as follows:

$$\left[\begin{array}{c}\mathbf{Y}_1\\ \hline \widetilde{\mathbf{Y}}_2\\ \widetilde{\mathbf{Y}}_3\end{array}\right]=\left[\begin{array}{cc}\mathbf{L}_{11} & \mathbf{0}\\ \hline \mathbf{L}_{21} & \mathbf{L}_{22}\end{array}\right]*\left[\begin{array}{c}\mathbf{L}_{11}^{-1}\mathbf{Y}_1\\ \hline \mathbf{W}\end{array}\right]$$

$$=\left[\begin{array}{c}\mathbf{Y}_1\\ \mathbf{L}_{21}\mathbf{L}_{11}^{-1}\mathbf{Y}_1+\mathbf{L}_{22}\mathbf{W}\end{array}\right] \tag{24}$$

where $\mathbf{W}$ is a random vector of (m) iid $N(0,1)$ random variables. Multiple realizations of the vectors $\widetilde{\mathbf{Y}}_2$ and $\widetilde{\mathbf{Y}}_3$ each conditional to $\mathbf{Y}_1$ (and to each other) are obtained by generating realizations of the *iid* random vector $\mathbf{W}$ and evaluating,

$$\begin{bmatrix}\widetilde{\mathbf{Y}}_2\\ \widetilde{\mathbf{Y}}_3\end{bmatrix}=\mathbf{L}_{21}\mathbf{L}_{11}^{-1}\mathbf{Y}_1+\mathbf{L}_{22}\mathbf{W} \tag{25}$$

for each realization of $\mathbf{W}$. A single conditional simulation of the SMU grade at location $\mathbf{u}_0$ is given by:

$$z_{\tilde{v}}(\mathbf{u}_0)=\frac{1}{t}\sum_{k=1}^{t}\varphi^{-1}(\widetilde{y_3}(\mathbf{u}_k)) \tag{26}$$

The corresponding estimated SMU grade made from conditionally simulated blast hole grades is given by:

$$z_{\tilde{v}**}(\mathbf{u}_0)=\sum_{j}^{s}\lambda_j B[\varphi^{-1}(\widetilde{y_2}(\mathbf{u}_j))\ ] \tag{27}$$

where λ are ordinary kriging weights for example and $\varphi^{-1}[\widetilde{\mathbf{Y}}_2]$ are simulated DDH values at the blast hole locations. $B[\cdot]$ is a user defined function for transforming simulated DDH grades to simulated BH grades. For example, the function $B[\cdot]$ could be used to add noise or deviations to the vector of simulated DDH grades $\varphi^{-1}[\widetilde{\mathbf{Y}}_2]$ (Parker and Isaaks, 1992; Journel and Kyriakidis, 2004).

The distributions $F_{\tilde{v}}(z;\mathbf{u}_0 \mid (n))$ and $F_{\tilde{v}}(z;\mathbf{u}_0 \mid \tilde{v}^{**},(n))$ of the conditionally simulated values $z_{\tilde{v}}(\mathbf{u}_0)$ and $z_{\tilde{v}^{**}}(\mathbf{u}_0)$ are generated by repeated applications of (25), (26), and (27). For example by using an efficient LU algorithm it may be practical to simulate as many as 500 equi-probable pairs of $z_{\tilde{v}}(\mathbf{u}_0)$ and $z_{\tilde{v}^{**}}(\mathbf{u}_0)$ for each SMU.

Thus, the simulated tonnage recovered over the domain D is given by:

$$\tilde{T}_D(z) = T_o[1 - F_D(z;\tilde{v}^{**} \mid (n))] \qquad \forall z \tag{28}$$

The simulated actual quantity of recovered metal is given by:

$$\tilde{Q}_D(z) = T_o \int_z^\infty z'\, dF_D(z';\tilde{v} \mid \tilde{v}^{**},(n)) \qquad \forall z \tag{29}$$

The simulated actual recovered grade is given by:

$$\tilde{m}_D(z) = \frac{\tilde{Q}_D(z)}{\tilde{T}_D(z)} \tag{30}$$

Equation (26) solves the change of support problem by computing a simple spatial average from a number of jointly simulated point values within the SMU. Note that each simulated point value is back-transformed before averaging.

Equation (27) provides a simulation of the grade control estimator using simulated blast hole grades. Note, that (27) includes a user-definable function enabling the user to simulate the relationship between the DDH and BH grades if known (Parker and Isaaks, 1992; Journel and Kyriakidis, 2004). Thus, the impact of poorer quality blast hole assays on the predicted recoveries can be put into the estimation of recoverable resources here. Equations (29) and (30) simulate the actual recovered quantity of metal and recovered grade given that the SMU are selected by their grade control estimate. The key is the simulated conditional distribution of the true SMU grades given their grade control estimates. This distribution quantitatively accounts for any conditional bias inherent in the grade control estimator as well as for any associated misclassification.

7 Conclusions

- If the block estimates are to be used for selection (grade control), then it is desirable to minimize conditional bias. Although conditional bias may be minimized, it likely cannot be eliminated.
- If the grade control estimator is conditionally biased, the predictions of the long-term mine planning model should quantitatively account for the bias. Such an accounting can be evaluated through conditional simulation.
- If the block estimates are not used for selection at the time of mining, but rather for the prediction of the tons and grade that will be recovered in the future, then whether or not the block estimator is conditionally biased is irrelevant to the accuracy of predicting the future recoveries.
- The predictions of the long-term model should quantitatively account for the ore loss and dilution (misclassification) that will occur at the time of mining. Again, such an accounting can be evaluated through conditional simulation.

- And finally, conditional simulation provides an easy solution to change of support. Long-term mine planning models with block sizes equivalent to the SMU are easily simulated. With good software, conditional simulation via the LU decomposition of the covariance matrix is as practical as ordinary kriging.

8 References

David, M., Marcotte, D. and Soulie, M., Conditional bias in kriging and a suggested correction, In Geostatistics for Natural Resource Characterization, G Verly, M David, AG Journel & A Maréchal, eds, Riedel Publishers, Dordrecht, pp 217-244, 1984.

David, M., *Geostatistical Ore Reserve Estimation*. Elsevier, Amsterdam, 1977.

Davis, M.W., Production of conditional simulations via the LU decomposition of the covariance matrix. Math Geology, 19(2):91-98, 1987.

Deraisme, J. and Roth, C., The information effect and estimating recoverable reserves. Geovariances. www.geovariances.fr/publications/article6/index.php3, 2000.

Deutsch, C.V. and McLennan, J.A., Conditional bias of geostatistical simulation for estimation of recoverable reserves. Can. Inst. Min. Metall. Bull., 2003.

Guertin, K., Correcting conditional bias, In Geostatistics for Natural Resource Characterization, G Verly, M David, AG Journel & A Maréchal, eds, Riedel Publishers, Dordrecht, pp 245-260, 1984.

Isaaks, E. H. and Srivastava, M. R., *An Introduction to Applied Geostatistics*. Oxford University Press, New York, 1989.

Journel, A.G. and Huijbregts, C.J., Mining Geostatistics. Academic Press, New York, 1978.

Krige, D.G., A basic perspective on the roles of classical statistics, data search routines, conditional biases, and information and smoothing effects in ore block valuations. Conference on Mining Geostatistics , Kruger National Park, 1994.

Krige, D.G., A practical analysis of the effects of spatial structure and of data available and accessed, on conditional biases in ordinary kriging. 5th International Geostatistics Congress, Wollongong, Australia, 1996.

Krige, D.G., Conditional Bias and Uncertainty of Estimation in Geostatistics. Keynote Address for APCOM'99 International Symposium, Colorado School of Mines, Golden, CO., 1999.

Matheron, G., The selectivity of the distributions and the "second principle of geostatistics", In *Geostatistics for Natural Resource Characterization, G Verly, M David, AG Journel & A Maréchal, eds*, Riedel Publishers, Dordrecht, pp 421 – 433, 1984.

Pan, G., Smoothing effect, conditional bias and recoverable reserves. Can. Inst. Min. Metall. Bull. V. 91, no. 1019, pp81-86, 1998.

Parker, H., The volume-variance relationship: a useful tool for mine planning. In P. Mousset-Jones, editor, *Geostatistics*, pages 61-69, New York, 1980. McGraw Hill, 1980.

Parker, H. and Isaaks, E.H, *The Assessment of Recoverable Reserves for the Minifie Deposit using Conditional Simulation*, Kennecott-Niugini Mining Joint Venture, Lihir Project Resource/Reserve Calculations, *1992.*

Sinclair, A.J., and Blackwell, G.H., *Applied Mineral Inventory Estimation*. Cambridge University Press, Cambridge, 2002.

POST PROCESSING OF SK ESTIMATORS AND SIMULATIONS FOR ASSESSMENT OF RECOVERABLE RESOURCES AND RESERVES FOR SOUTH AFRICAN GOLD MINES.

D.G. KRIGE [1], W. ASSIBEY-BONSU [2] AND L. TOLMAY [2]
[1] *Private Consultant, South Africa*
[2] *Gold Fields Limited, South Africa*

ABSTRACT. This study is based on a comprehensive data base from a section of a large deep South African gold mine. The upper section of the area covered was accepted as providing the known data for purposes of estimating in a deeper extension of this section. Ore blocks in this extension were valued using the data from the upper 'known' area together with data in development raises typically 150m apart in the deeper extension area. Estimation techniques used were Simple Kriging (SK) with post-processing, and Simulations and the recoverable block estimates were compared with the known follow-up 'actual' values of these blocks. The study shows that the direct SK post-processing and repeated simulation approaches, if applied efficiently, can provide equally useful tools for computing global recoverable resources. However, the direct SK post-processed technique provided the only advanced practical estimates of individual ore blocks for short-term mine planning, grade control and ore resource/reserve classification.

1 Brief Historical Background to Ore Block Valuations

The main objectives of block valuations in South African gold mines have always been, and still are:

- To provide management and shareholders with a reliable inventory of the mine's basic asset, i.e. its ore resources and reserves classified into categories as required by the relevant codes.
- The estimation of tonnages and grades expected to be obtained from mining in short and medium term time categories e.g. monthly, quarterly and annually, *and from individual stopes and mine sections.*
- Where the average ore grade is not sufficiently high to warrant 100% mining of the ore body, proper advanced indications for the selection of blocks above the break even or cut off.
- The planning of grade control so as to produce a profile of acceptable production and financial targets.

The birth of Geostatistics and kriging in South Africa more than 50 years ago resulted from the statistical explanation of the presence of conditional biases in the orthodox valuation techniques (Krige 1951). Kriging, properly applied, eliminated these biases

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 375-386.

but provided smoothed estimates. To overcome the smoothing effect, some geostatisticians introduced the practice of Ordinary Kriging (OK), but with a limited data search, and without using the global mean or applying simple kriging (SK). This could rectify the problem of smoothing but had the effect of re-introducing the conditional biases, which in fact had led to the birth of geostatistics. The practice of using a limited search cannot be condoned (Krige, 1996, 1997, 2001; McLennan and Deutsch, 2004). It is also theoretically impossible to meet both objectives of conditional unbiasedness and the absence of smoothing *on the basis of specific fixed grade estimates for individual blocks.*

A logical advance towards a solution of the problem of smoothing was the substitution of probability estimates for fixed individual kriged estimates. This was effected by using uniform conditioning, direct or indirect conditioning (Assibey-Bonsu and Krige 1999b) and various other post processing procedures, e.g. spectral postprocessor (Journel, Kyriakidis and Mao 2000). Simulation techniques have also been proposed for producing unsmoothed and unbiased block recoverable estimates. However, single simulations could be unsmoothed but will be conditionally biased (Krige and Assibey-Bonsu, 1999a), and repeated simulations, when averaged, will produce smoothed values. Lately, McLennan and Deutsch (2004) have suggested "conditional non-biased simulation" based effectively on the introduction of the concept of probability estimates via repeated simulations, in substitution for specific block estimates. This is a form of post processing, or conditioning, as practiced in the direct processing of kriged estimates.

The authors (McLennan and Deutsch (2004)) compared such estimates with straight kriged estimates, but not with kriged estimates after post-processing. Their analyses covered only estimates on a global basis. The comparison of these techniques for estimates for individual blocks and local small production areas, and the overall effect on grade profiles over time, were not considered.

The above background calls for block estimates to be *globally unbiased as well as for individual blocks and mine sections*, and also properly processed to eliminate any 'smoothing' effects. The argument that final selection of ore blocks as ore or waste is done at the stage when the more intensive sampling data are available on a proper SMU (Selective Mining Unit) basis and that 'unbiasedness' and 'unsmoothing' are only necessary on a global basis, does not hold except possibly, but still to a more limited extent only, for open cast mines. A detailed examination and comparison of kriged estimates before and after post processing with the recent simulation approaches on both a global and more local short-term basis, including individual blocks, is therefore justified. This is the objective of this paper based on a massive set of 'actual' data from a large deep level South African gold mine.

2 Basic Principles of the Two Main Techniques

Conditioning of unbiased estimates is based on the principle of replacing these smoothed estimates with probability distributions representing the expected follow-up 'actual' grades with each kriged estimate as the mean of the distribution at a variance level equal to that expected for the 'actual' grades. For *direct conditioning* this is effected by super imposing on the kriged estimate for each block, a 'simulated' distribution of expected 'actual' values with a variance equal to the difference in variability between the smoothed and 'actual' grades (Assibey-Bonsu and Krige, 1999b). The end result is an estimated unsmoothed tonnage-grade curve to replace the smoothed kriged estimate. For *uniform conditioning* this is done, not for individual SMU blocks, but for groups of local SMU blocks within larger panels or blocks. In this study all references to SMU's and 'blocks' refer to 2D ore units of 20x20m in the plane of the ore body.

The Sequential Gaussian Simulation (SGS) technique was used for generating the simulation realizations. Simple Kriging was used to determine the parameters of the Gaussian conditional cumulative distribution function at respective locations. The SGS was generated using the GSLIB software (See Clayton and Journel, 1992). Post-processing of the simulation results adopted in this paper is similar to that proposed by McLennan and Deutsch (2004). It involves the distribution of the repeated simulated values, say 50, for each block as an estimate of the unsmoothed tonnage-grade curve for the block, and thus that it reflects the uncertainty of the mean of the relevant 50 realisations as an estimate of the 'actual' grade. The assumption is, thus, that the variance of the 50 simulated values straddling the mean of the 50 simulated grades for each block reflects the probability distribution of the 'actual' grades.

The objective of this analysis is to apply these two main techniques using a set of 'actual' data to provide estimates for comparison with 'actual' follow-up information on a practical basis, so as to determine the validities of the alternative techniques and their relative efficiencies.

3 The Data Base

The 'actual' data used should ideally represent the grades of SMU blocks as estimated on the basis of the more extensive data, which will be available at the time of the actual selection during production. Such data for a block can never be complete and *thus the presence of the inevitable information effect in the setting of the target of the SMU 'actuals' for the dispersion variance of block estimates and of the tonnage-grade curves is an essential requirement for both the SK post-processing and the repeated simulation techniques.*

The data used in this study comprised an area of 2Km x 2Km on the Ventersdorp Contact Reef (VCR) on a large deep mine with a total number of nearly 43000 underground sample values recorded as cm.g/t grades and reflecting a direct measure of the gold concentration per unit area of the ore body. The ore body strikes approximately in a north south direction at an average dip westwards of 28^0, and with an average mining width of 139 cm.

Fig. 1 shows the total area covered. The VCR in this area forms a geologically homogenous population with no significant grade trend with depth. The area was split into an upper section of 2Km x 1Km with about half of the samples from stope (or panel) faces and from development exposures, including a set of 8 raises 150m apart and extending into strips 1and 2 in the lower follow-up area. These values are accepted as 'point' values from the 'known' database to be used to estimate ore resource blocks of 20m x 20m in the deeper 'follow-up' section. The valuation is largely an extrapolation exercise since no regular underground drilling is practiced (Assibey-Bonsu and Krige, 2003).

The 'known' point values in the deeper section were also regularized into 363 data blocks and used as the follow-up 'real' block values for judging the comparative efficiencies of the estimates for these blocks. Follow-up blocks with less than 5 samples per block were discarded because the information effect for such blocks will be abnormally high. The 'known' block values for both the upper and lower areas are shown on Fig. 1 in 4 shades of grey/black and the patterns for both areas confirm the absence of any significant trend.

The geostatistical details of the 'actual' data for points and blocks are recorded in Table 1, and the 3-parameter lognormal distribution model and variogram in Figs 2 and 2B respectively. The third parameter of 255 cm g/t provides acceptable fits for the lognormal distributions of points and of block grades. Table 1B shows the variogram parameters for normal score and relative models. Fig. 4 also shows the target tonnage grade curves for the 363 'actual' 20 x 20m blocks for the follow-up area.

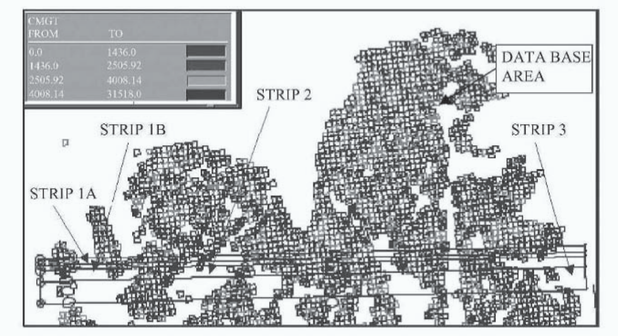


Figure 1: Showing study area and strips representing different production periods.

	No Units	Mean cm.g/t	Std Dev	3 Par Log n b=255	
				Mean	Variance
Upper Area					
Points	18684	3061	4660	7.541	1.15
Blocks 20 by 20m	1335	2918	2317	7.907	0.403
Follow up area					
Total Points	23372	3281	4210	7.667	1.064
Strips 1 to 3:					
Blocks 20 by 20m	363	3311	2420	7.961	0.466

Table 1: Showing details of the database of point and 20x20m block grades

	DIRECTION	NUGGET	C1	R 1	C2	R 2	C3	R 3
POINTS	120 DEGREES	0.414	0.372	30	0.116	70	0.098	500
	30 DEGREES			25		70		180
BLOCKS	120 DEGREES	0.26	0.4	65	0.3	85	0.03	702
20 BY 20	30 DEGREES			32		85		204
Relative	30 DEGREES	0.126	0.153	33	0.14	85	0.068	204
20 BY 20	120 DEGREES			43		63		700

Table 1B: Showing relative and normal score semi variogram parameters.

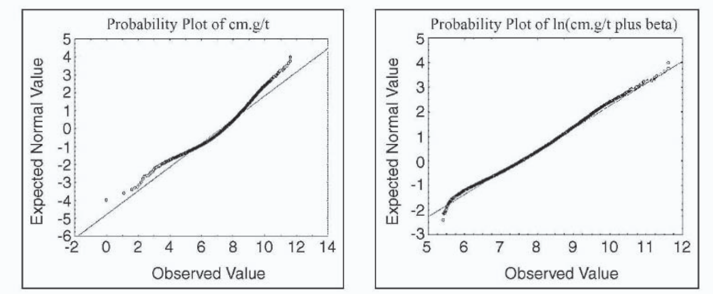


Figure 2**:** Showing 3-parameter lognormal distribution models with additive constants of zero and 255 for Data Base (Points).

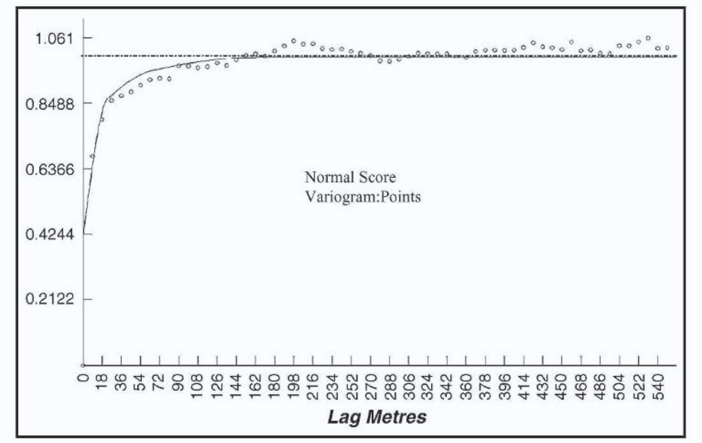


Figure 2B: Showing Normal Score variogram in 30 degrees direction for points.

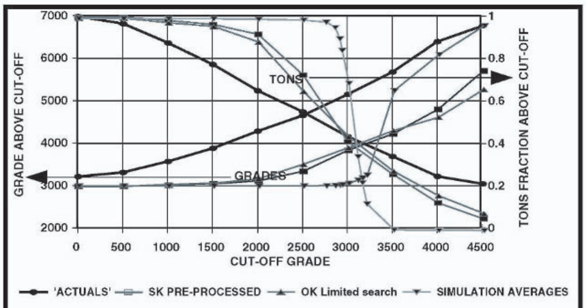


FIG. 3 SHOWING THE GLOBAL TONNAGE-GRADE CURVES FOR 'ACTUALS', SK PRE-PROCESSED, OK LIMITED SEARCH and SIMULATION AVERAGES

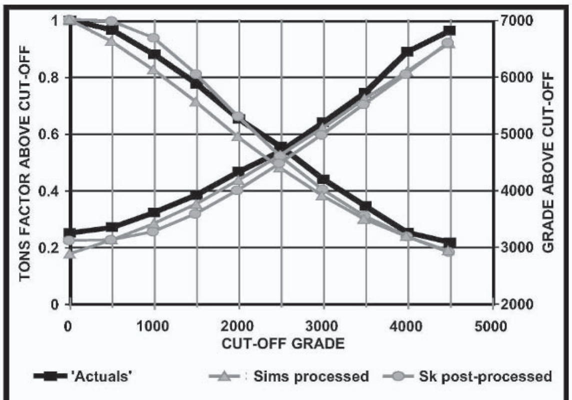


FIG.4 showing the global tonnage-grade curves for the total follow-up area for 363 blocks for: 'Actuals', SK processed and Repeated Simulations.

4 Estimation Techniques – Global Results

For all comparisons of dispersion variances, tonnage-grade curves, and correlations with slopes of regression, etc., the estimates and 'actuals' were normalized by transforming the grades to Logarithm (grade + 255) (see Fig. 2). This caters for the proportional effect, and ensures the approximation to Normal distributions, and thus provides for linear regression trends with slopes for measuring the presence and extent of conditional biases.

The following techniques have been used for the global follow-up area

- *Ordinary Kriging with a limited search.*

This technique is aimed at overcoming the 'smoothing' effect of an extensive search routine or of Simple Kriging (SK). The search parameters used were as follows:

Minimum number of point data…………………...2
Maximum number of point data…………………..8

The results and comparisons with the follow-up 'actual' data are summarized in Table 2 and shown in Fig.3. *The estimates show some elimination of the smoothing effect but serious conditional biases and cannot be recommended for detailed mine planning purposes.*

		Dispersion Variances		OK vs Act	
STRIP	No.Blks	Actuals	OK	Corr.Coef.	Regr.Slope
1/3 total	363	0.466	0.204	0.401	0.606

Table 2: Showing dispersion variances and correlation details for 'actuals' and OK estimates on Ln(x+b) basis.

- *Simple Kriging with no post processing.*

With no post-processing or conditioning, this technique largely overcomes the problem of conditional biases (slope of regression = 0.96) but the results are 'smoothed' with a Ln Variance of 0.087 compared to that for the 'actuals' of 0.47 (see Fig 3, Table 3 and

4). The results provide some provisional indications for the selection of blocks above cut-off but will be conservative for recoverable grades and optimistic for the corresponding tonnages. *Nevertheless, this technique provides a base for performing post-processing for which the effective absence of conditional biases is essential and it provides some advanced indication of which blocks are likely to be mined above cut-off.*

- *Simple Kriging with post-processing*

The above results were post-processed with direct conditioning (Assibey-Bonsu and Krige, 1999b) to provide the comparison with 'actuals' on a tonnage grade basis as shown in Fig. 4 and Table 3. For this purpose ten cut-off grades were used as shown in the figure. The dispersion variance of the estimates and the grade-tonnage curve are not ideal but approach those of the 'actuals'.

		Ln Variances			Variances of Probability distributions		
STRIP	No. Blks	Actuals	SK Pre.	Sim Avgs	SK proc.*	Sims.direct	Equiv. sims
1A total	59	0.318	0.111	0.041	0.310	0.597	0.441
1Btotal	58	0.379	0.099	0.022		0.556	0.435
1A+1B	117	0.348	0.105	0.032		0.577	0.438
2	96	0.479	0.115	0.021	0.360	0.563	0.422
3	150	0.536	0.054	0.012		0.556	0.433
1+2 total	213	0.414	0.110	0.027			0.431
1/3 total	363	0.466	0.087	0.021			0.432

Table 3: Showing dispersion variances for 'actual', and SK and Sims. Pre- and post-processed estimates on Ln(x+b) basis.
* Graphical

The differences between these estimates and the 'actuals' result from dispersion variances for the SK estimates of 0.31 to 0.36 compared to the follow-up variance of 0.35 (strip 1, i.e. 1A+1B) to about 0.5 (strips 2 and 3), i.e. an apparent remaining smoothing effect. However, the 'actual' dispersion variance is too high due to the presence of a low information effect resulting from an average of some 12 values inside each follow-up block. The SK estimates cover a higher information effect and a correspondingly lower dispersion variance effectively in line with the actual position during production when selections are restricted to values external to the blocks. *This stresses the importance of all post-processing procedures to take proper account of a realistic information effect.*

	Correlation Coef . with 'actuals'			Regr. Slope		
STRIP	S Kpre	SimAvgs	first simn	SKpre	SimAvgs	first simn
1A total	0.645	-0.113	-0.147	1.108	-0.315	-0.100
1Btotal	0.436	-0.111	0.282	0.854	-0.458	0.242
1A+1B	0.548	-0.110	0.061	1.034	-0.364	0.046
2	0.309	-0.046	-0.048	0.644	-0.223	-0.039
3	0.380	-0.022	0.050	1.380	-0.143	0.046
1+2 total	0.431	-0.077	0.010	0.860	-0.302	0.008
1/3 total	0.392	-0.057	0.028	0.956	-0.269	0.023

Table 4: Showing correlations and regression slopes for SK and Sims vs. 'actuals' on Ln(x+b) basis.

- *Repeated simulations – block averages for 50 iterations*

The simulation approach recently proposed by McLennan and Deutsch (2004) was applied to the point data from the database and using the corresponding semi-variogram. The results from 50 iterations were first averaged for each block to provide the 'smoothed' block tonnage grade curve shown on Fig. 3 and the correlation results in Tables 3 and 4. *Unlike the original SK un-processed block estimates, these simulation averages do not meet the requirement for the effective absence of conditional biases.*

- *Repeated simulations on a probability basis.*

In this approach the 50 simulated grades for each set of iterations are accepted as a probability model for each block. The resultant global tonnage grade curve and results, shown in Fig 4 and in Table 3, serve the purpose of comparison with the 'actuals' and SK with post-processing above. In this case the dispersion variance for the estimates of 0.43 compares well with that of 0.47 for the 'actuals'. However, a study of the sets of simulation distributions show a departure from the 3-parameter model used for the other block estimates and 'actuals'. To overcome this problem, the untransformed mean and variance for each block were used to calculate the theoretical equivalent 3-parameter variances for the simulation distributions with the same untransformed means and variances. These averaged 0.432 for strips 1 to 3 which agrees reasonably well with the 'actuals'. Note that *the information effect did not feature in the simulation process, and effectively produces results with a variance close to that for perfect block valuations, i.e., the grades indicated can be too optimistic*

5. Main Conclusions For Global Estimates

The techniques covered in paragraph 4 above, other than SK processed and Simulations on a probability basis, cannot be recommended and leave only the latter two for further consideration.

Note that the repeated simulations do little to distinguish between individual blocks for the guiding of the selection process in mine planning in advance of actual selection during production. At the other extreme end the position for a single simulation for individual blocks shows virtually no correlation with the 'actuals' and serious conditional biases (coefficient = 0.028, and regression slope = 0.023, see Table 4, first simulation). *The latter highlights the danger of selection of any single simulation realization, e.g., median realization, for mine planning.*

5.1 RELATIVE ECONOMIC PROFITS

The two techniques have also been compared with the actuals on the basis of an elementary financial analysis of *'relative profits'* defined as:

(tons above cut-off) x (grade above cut-off – cut-off)

and the results are shown in Fig 5 for the global position for the follow-up area. There is a reasonable agreement of both approaches with the 'actual'. *The critical conclusion is that globally the two approaches can produce results very close to those for the follow-ups. The situation for individual blocks and local small areas will be discussed in the following paragraph.*

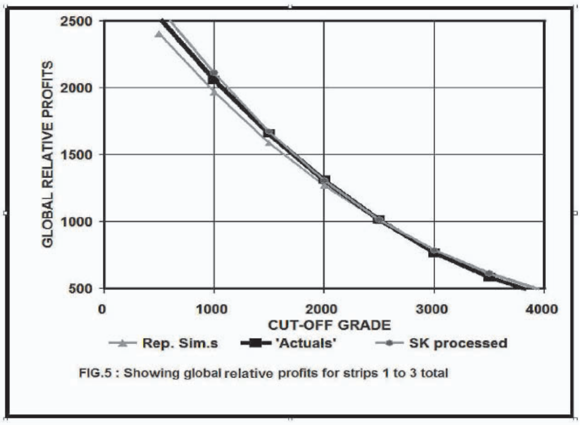


FIG.5 : Showing global relative profits for strips 1 to 3 total

6. Results for Subdivisons of the Global Area

Paragraphs 3 and 4 cover the global position for the whole follow-up area. In order to focus on the estimates and grade control problems specifically for short-term production, the follow-up area was split into 3 main strips as shown in Fig. 1. Strips 1A and 1B cover the first and second 20m extensions into the estimation area, with measured resources and mining periods of approximately 4 months each. Strips 2 and 3 cover further extensions of 40m and 80m respectively, both with indicated resources and with mining periods of 8 months to 16months and 16 to 32 months respectively.

The results for the 'actuals' and the 2 techniques remaining for further consideration have been further examined for the sub-divisions represented by the individual strips 1 to 3 (See Tables 3 and 4 and Fig. 4). The individual strips shows fairly stable dispersion variances for 'actuals', simulation averages and SK pre processed, but a decline for the latter two in their correlation levels with the 'actuals' (see Tables 3 and 4). This is to be expected as the distance of known data accessed for block valuations increases steadily from strip 1 to strip 3. *The regression slopes are heavily in favour of the SK pre-processed estimates* against negative slopes for Simulation averages (-0.3 to -0.46, i.e. serious conditional biases).

The post-processed versions of these two sets of estimates cannot be directly correlated with the 'actuals' in the light of their probability nature as distinct from the specific SK pre-processed and the simulation average figures for individual blocks and sections. However, the general tenor of the latter figures should carry through to the 'post-processed' versions.

For this reason the position for strip 1A was analysed in some detail and is summarized in Figs. 6A and 6B. The note under par. 5 above applies particularly to this strip when the two figures are compared for simulation averages and SK pre-processed estimates vs. 'actuals'. *Fig 6A demonstrates clearly that the simulation averages per block provide effectively no correlation with the 'actuals' and show maximum conditional biases; i.e. no contribution to the problem of doing selective planning on a block basis in advance of the final selection when more data will be available. In contrast to simulations the SK pre-processed results in Fig. 6B show a reasonable correlation level of 0.65 and virtually no conditional biases.* The general tenor of these figures in fact does carry through to the 'post-processed' versions, and is confirmed by the relative profits for these 2 techniques vs. 'actuals' as demonstrated in the correlation graphs in Fig 7A and 7B.

The results for the simulation averages and the corresponding probability estimates are evidently due to the fact that on the South African gold mines virtually no conditioning data are available for block estimates as these estimates are done essentially on an extrapolation basis. *The simulations are thus not fully conditional and should only be used for global patterns and tonnage/grade curves*.

7. Overall Conclusions

This paper stresses the main principles in geostatistical applications to mine resource valuation, which should be accepted and practiced by all concerned:

i) Where at all practical, alternative techniques and their detailed procedures followed, such as various kriging approaches with a choice of search routines, simulations etc., should be compared using an actual data base. This will provide actual follow-up data for correlations with the estimates, including measures of comparable efficiencies. Where such actual data are not available, a suitable simulation base could be used. It is disturbing that the geostatistical literature over many years of outstanding achievements, have provided very few such studies. They could have eliminated many misunderstanding between some practitioners.

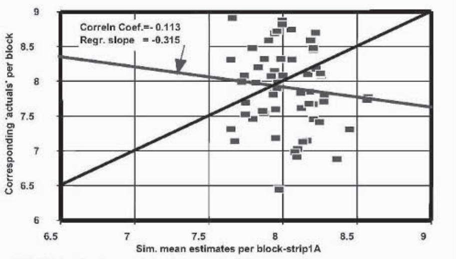


FIG. 6A showing the correlation between simulation averages per block and the corresponding ,actuals' on Ln(x +255) basis-strip 1A.

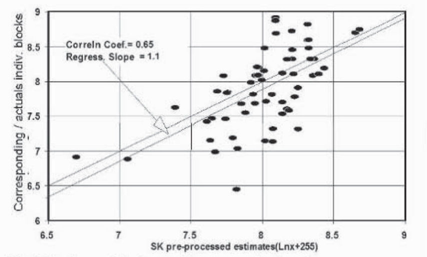


FIG. 6B showing correlation between SK pre-processed estimates for individual blocks in Strip 1Avs corresponding 'actuals' on Ln(x+255)

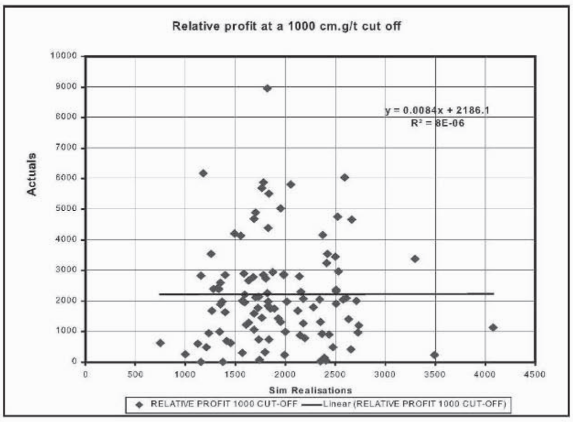


Figure 7A: Relative Profit at a 1,000cmg/t cut-off (simulation)

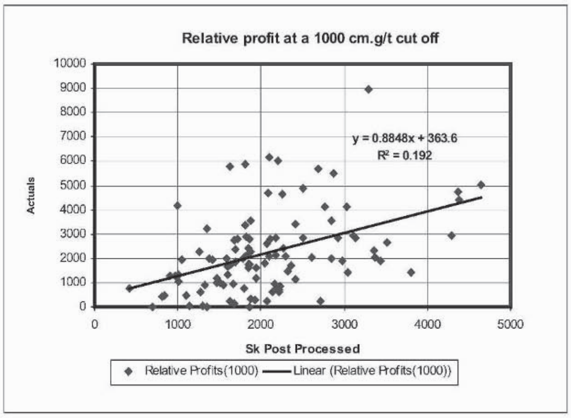


Figure 7B: Relative Profit at a 1,000cmg/t cut-off (SK post-processing)

ii) The principle of conditional unbiasedness, which gave rise to the birth of geostatistics more than fifty years ago, is still valid today. This principle cannot be reconciled with any unsmoothed estimates such as OK with limited search routine. Any new technique, however sophisticated, must be tested on a practical follow-up study, as mentioned above. The only solution at this stage is via some form of probability estimates, as used in this study.

This paper shows that, for global estimates, there is little to choose between SK post-processed and simulation probability estimates, particularly for deep level mining where blocks are valued largely on an extrapolation basis. For short-term individual block estimates, however, kriging with post-processing shows a distinctive advantage over repeated simulations. For any mine, where some advanced drilling is available at the resource valuation stage, a detailed practical study, similar to this, seems necessary to compare the alternative techniques at various levels of data densities, particularly for individual short term block estimates. However, an earlier pilot study (Assibey-Bonsu and Krige, 1999a), but not on the same scale and basis, indicated similar results as shown in this paper.

8. Acknowledgements

Acknowledgement is made to Gold Fields Limited, for permission to publish this paper.

9. References

Assibey-Bonsu, W. and Krige, D.G. (1999a) Practical *Problems in the estimation of recoverable reserves when using Kriging or Simulation Techniques*. International Symposium on Geostatistical Simulation in Mining, Perth Australia, October 1999

Assibey-Bonsu, W. and Krige, D.G. (1999b) Use *of Direct and Indirect Distributions of Selective Mining Units for Estimation of Recoverable Resource/Reserves for New Mining Projects*, APCOM'99 International Symposium, Colorado School of Mines, Golden, October 1999.

Assibey-Bonsu, W. and Krige, D.G.(2003). *An analysis of the practical and economic implications of systematic underground drilling in deep South African gold mines.* APCOM 2003 International Symposium, Cape Town, May 2003 (SAIMM).

Journel A.G., Kyriadkidis P.C., and Mao S. (2000). *Correcting the smoothing effect of estimators: a spectral postprocessor*. Mathematical geology, Vol. 32, No7, October 2000.

Deutsch C. V. and Journel A.G. (1992). *Geostatistical Software Library and User Guide.* Oxford University Press, 1992., pp340

Krige, D.G. (1951). *A statistical approach to some basic mine valuation problems on the Witwatersrand :* J. of the Chem. Metall. and Min. Soc. of S.A. December, 1951 - discussions and replies March, May, July and August 1952.

Krige, D.G. (1960). *On the departure of ore value distributions from the log-normal model in South African gold mines*. J.S.A.I.M.M., November 1960, January and August 1961.

Krige, D.G. (1962). *The application of correlation and regression techniques in the selective mining of gold ores*. 2nd APCOM Symposium, University of Arizona, April, 1962.

Krige, D.G. (1996). *A practical analysis of the effects of spatial structure and data available and used, on conditional biases in ordinary kriging* - 5th International Geostatistics Congress, Wollongong, Australia, 1996.

Krige, D.G. (1997). *Block Kriging and the fallacy of endeavouring to reduce or eliminate smoothing. Keynote address*, 2nd Regional APCOM Symposium, Moscow State Mining University, August 1997

Krige, D.G. (2001). *Comment on paper by Journel and others on a Spectral Postprocessor*. Mathematical Geology,Vol 33, No.6, 2001.

McLennan, J.A. and Deutsch, C.V. (2004*) Conditional Non-Bias of Geostatistical Simulation For Estimation of Recoverable Reserves*. Canadian Inst. of Min. and Met. (CIM) Bulletin, May 2004.

THE PRACTICE OF SEQUENTIAL GAUSSIAN SIMULATION

MAREK NOWAK[1] and GEORGES VERLY[2]
[1] Nowak Consultants Inc. 1307 Brunette Ave, Coquitlam BC V3K 1G6
[2] Placer Dome Inc. 1055 Dunsmuir St, Vancouver BC V7X 1P1

Abstract. The theory of simulation is relatively well documented but not its practice, which is a problem since simulation is not as robust as linear estimation. As a result, many costly mistakes probably go undetected. In this paper, a process for simulation is introduced with the objective of reducing the likelihood of such mistakes. The context is sequential Gaussian simulation within the mining industry. However, a significant part of the process can be applied in other simulation framework.

Four of the most important aspects of the process are discussed in detail. A gradual trend adjustment is suggested as a post-simulation step. A modified bootstrap approach is presented to deal with the grade uncertainty that accounts for spatial dependence between the samples. A number of pre- and post-simulation checks are also discussed. Some post-simulation adjustments of the simulated values are suggested to improve on the quality of the simulation.

All of the approaches, solutions and checks presented in this paper are simple, flexible, and can be easily implemented by a practitioner.

1 Introduction

Sequential Gaussian simulation starts by defining the univariate distribution of values, e.g., assay grade values, performing a normal score transform of the original values to a standard normal distribution, and assuming multi-normality of the normal scores. The multi-normal assumption ensures that the conditional distribution at a given location is normal with mean and variance provided by simple kriging (SK). Simulation of normal scores at grid node locations is done sequentially most often with SK using the normal score variogram and a zero mean (Isaaks, 1991, Deutsch and Journel, 1998, Goovaerts, 1997). Once all normal scores are simulated, they are back-transformed to original grade values.

Although the simulation methodology is well documented, a practical process leading to valid and representative realizations of in-situ grades is rarely a focus of attention within the geostatistical community. To a practitioner, this can lead to frustration in applying a methodology that may produce poor results. There is a need for a simulation procedure that is systematic, robust and easy to follow. This need led Placer Dome to design a

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 387-398.

process for sequential Gaussian simulation. Note that a significant portion of the process can be applied to other simulation algorithms (see Figure 1).

Figure 1 shows that the process is more complex than just normal score transformation, variogram modeling, simulation and back-transformation. A number of steps have been added to improve on the span of uncertainties, trend reproduction, reproduction of data distribution, reproduction of a variogram model, reproduction of correlated variables, and choice of optimistic and pessimistic scenarios. Although some of these steps have yet to be implemented, generally the process is closely followed by Placer and is described in Nowak and Verly (2004).

This process is based on specific difficulties encountered during real case studies, in particular:

- Simulated values may not adequately follow general trends, especially away from data locations.
- Bootstrapped distributions may be almost identical when created from large data sets.
- Average and/or variability of simulated data may be substantially different from average and/or variability of the conditioning data.
- Variograms of simulated values may be different from the variogram models.

This paper is a detailed discussion of the following portion of the process:

- Trend analysis.
- Bootstrap grades.
- Check/Adjust simulated normal scores (histograms and variograms).
- Check/Adjust distribution of simulated grades.

2 Trend analysis

Trends are not always well reproduced in sequential Gaussian simulation (Steps I.1.2 and I.1.8 in Figure 1a). This is because of the stationarity assumption necessary for the normal score transform and the assumption of a constant zero mean in the SK algorithm. One simple way to deal with this problem is to filter the trend and simulate the residuals of the original values (Deutsch, 2002). Unfortunately, this solution may produce simulated grade values that are negative. An obvious way out is to reset the negative values to zero, but this may result in a significant bias and poor reproduction of the trend.

A second solution consists in defining the local prior means to be used by SK with a correction factor for all kriging variances (Goovaerts, 1997; Deutsch, 1998). This solution was not tried by the authors but it is suspected that it may lead to difficulties in the reproduction of the original values distribution and it does not address the fact that the normal score transform is global within a geological domain.

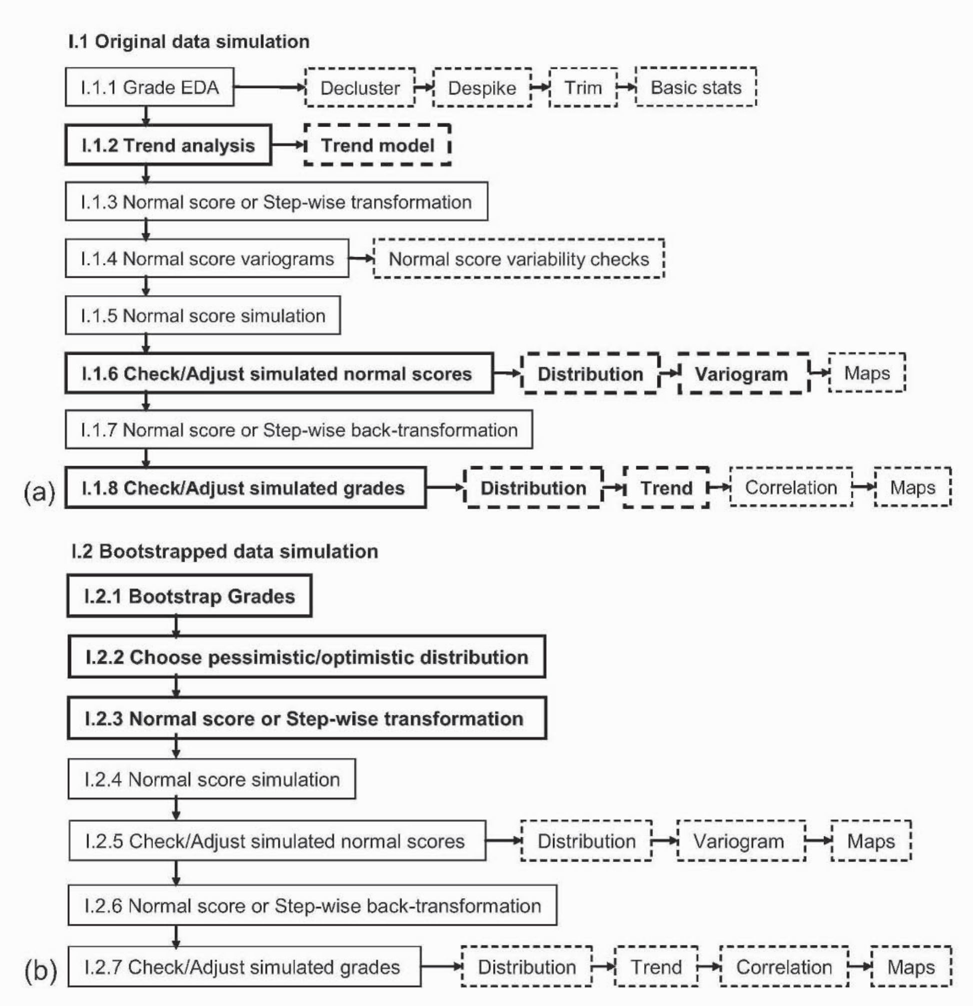

I.1 Original data simulation
I.1.1 Grade EDA
Decluster
Despike
Trim
Basic stats
I.1.2 Trend analysis
Trend model
I.1.3 Normal score or Step-wise transformation
I.1.4 Normal score variograms
Normal score variability checks
I.1.5 Normal score simulation
I.1.6 Check/Adjust simulated normal scores
Distribution
Variogram
Maps
I.1.7 Normal score or Step-wise back-transformation
(a)
I.1.8 Check/Adjust simulated grades
Distribution
Trend
Correlation
Maps
I.2 Bootstrapped data simulation
I.2.1 Bootstrap Grades
I.2.2 Choose pessimistic/optimistic distribution
I.2.3 Normal score or Step-wise transformation
I.2.4 Normal score simulation
I.2.5 Check/Adjust simulated normal scores
Distribution
Variogram
Maps
I.2.6 Normal score or Step-wise back-transformation
(b)
I.2.7 Check/Adjust simulated grades
Distribution
Trend
Correlation
Maps

Figure 1. Process for simulating (a) original data and (b) bootstrapped data. The topics discussed in this paper are highlighted. The other process steps are discussed in Nowak and Verly (2004).

A third solution is given by Leuangthong and Deutsch (2004) who suggest a step-wise normal score transform. The method consists in defining the trend and residuals followed by a normal score transform of the residuals conditional to the trend. In practice, the residuals are classified according to a series of trend value intervals and there is one standard normal score transform of the residual per interval. This method is very promising because the normal score transform is conditional to the trend. The method ensures that there is no trend in the normal score space and that a proper normal score variogram is used. Finally, the method greatly reduces the number of negative grade values after the step-wise back-transform.

This method can be modified to a transformation of the original values conditional to the trend instead of residuals conditional to the trend, which would ensure that there are no negative grades after back-transformation.

Although the step-wise normal score transform is very promising, other solutions for trend reproduction have been tried by the authors. These solutions rely on a definition of a trend at all grid locations and on the average simulated model. It is assumed that the trend represents a relatively smooth surface and can be assessed by OK with a high nugget effect. An example of the trend values compared with the original data is presented in Figure 2a. The first attempt consisted of filtering the trend and simulating residuals. This approach, however, was abandoned because of a significant amount of simulated negative grades. Other attempts were made to correct for the trend of the simulated normal score values or the back-transformed simulated values. The best results have been obtained by adjusting back-transformed values according to:

$$Sim_{tr}(x) = Sim(x) \cdot w(x) \tag{1}$$

where *Sim(x)* is the simulated value at location x before the trend adjustment**,** $Sim_{tr}(x)$ is the simulated value after trend adjustment**,** and *w(x)* is a correction factor calculated as follows:

$$\begin{aligned} w(x) &= (c(x) - 1) \cdot v(x) + 1 \\ c(x) &= Tr(x) / Avsim(x) \\ v(x) &= \sigma_{kr}(x) / \sigma_{kr\,\max} \end{aligned}$$

where *Tr(x)* is the trend value at location x, *Avsim(x)* is the average simulated value, $\sigma_{kr}(x)$ is the kriging standard deviation and $\sigma_{k\ \mathrm{rmax}}$ is the maximum kriging standard deviation at any given node.

The kriging standard deviation $\sigma_{kr}(x)$ affects the amount of the adjustment. If a simulated node is very close to conditioning data then $v(x) \cong 0$ and no adjustment is made. On the other hand, a maximum adjustment is made far from data locations. Note that a similar progressive correction, i.e., a correction dependent on the distance from the data, has been discussed by Xu (1997). The advantages of the approach are:

- The average of simulated values is similar to the trend, in particular away from data locations.
- The coefficients of variation of the simulated values before and after the correction have been observed to be quite similar in practice.
- The correction is simple and can be done on already simulated values.
- The correction is flexible in the sense that $\sigma_{k\ \mathrm{rmax}}$ can be replaced by an arbitrary value.

The disadvantage of the approach is the difficulty to infer the trend everywhere, in particular far from data locations.

Figure 2b shows a comparison of the trend with the average simulation before the trend adjustment, and Figure 2c presents the comparison after the adjustment for the trend. Clearly, there is a substantial improvement in the reproduction of the trend when the adjustment is made.

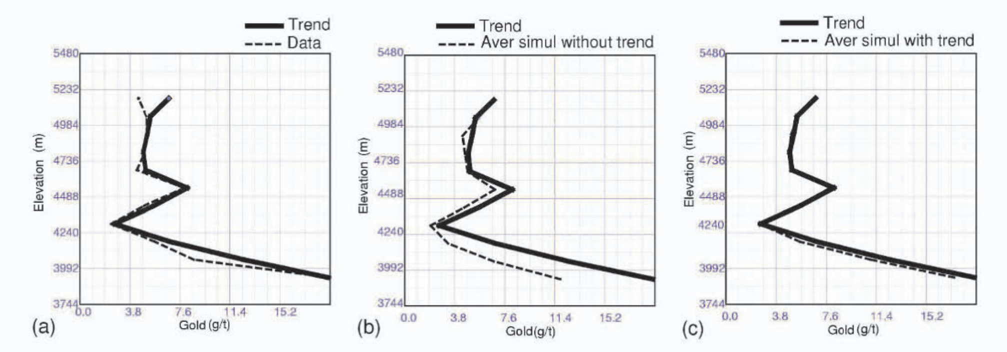


Figure 2. Comparison of the trend (solid line) along elevation with (a) conditioning data, (b) average simulation before trend adjustment, (c) average simulation after trend adjustment

3 Check/Adjust simulated normal scores

Post-simulation checks are necessary to ensure a reasonable reproduction of the distribution and spatial correlation (Step I.1.6 in Figure 1a). In a first step both the histogram and the variogram of the simulated normal scores are checked against the original normal score histogram and variogram. All the realizations should be considered at the same time for the checks to avoid natural fluctuations between the realizations. The verification of the results should take place within the same zone that has been used to get the simulation parameters, i.e., the declustered grade distribution, the normal score transform, and the normal score variograms (Figure 1a, Steps I.1.1, I.1.3, and I.1.4).

3.1 HISTOGRAMS

The simulated normal score histogram check may reveal that the simulated distribution is not standard normal. This section discusses (1) the case of the average of the simulated values different from 0.0, (2) the case of the variance of simulated values different from 1.0, and (3) a gradual adjustment of the simulated value to a standard normal distribution.

3.1.1 Simulated normal score average different from 0.0.

The difference may result from an improperly defined validation zone, i.e., the zone within which the simulation results are validated. Usually, this zone should be similar to the zone within which the simulation parameters (histogram, normal score transform, variogram) are calibrated. The difference may also result from an improper declustering of the original distribution. Two possible solutions are:

- Modification of the validation zone. If for example the non-zero average is due to a significant amount of simulated values at some distance from the conditioning data, and at the same time conditioned to low assays on the edges of the drilled out area, a modification of the validation zone that excludes areas far from conditioning data may reduce significantly the difference observed. Figure 3 illustrates the impact of such a modification of the validation zone. Here, in the original validation zone the average of simulated values is -0.11 but in the modified validation zone the average is -0.02 which is close to the 0.0 data average. The modified validation zone is limited to the area close to the conditioning data, extending not further than a search radius used for polygonal declustering.
- Adjustment to declustering weights. If a polygonal declustering is used, the search radius may be inappropriate. In other words, the original data distribution has not been properly defined.

If the source of the difference is unknown and there is reason to believe that the original distribution mean (= 0.0) is correct, the simulated values may have to be adjusted as per sub-section 3.1.3.

3.1.2 Variance of simulated values is different from 1.0

As for the average, the difference in variance may result from an improper validation zone or improper declustering and the solutions proposed earlier for correcting the average may be applied.

Another reason for a difference in variance is a possible inconsistency between the normal score transform and the normal score variogram. By construction, the normal score conditioning values are standard normal within the zone of interest *Z* (e.g. one geology domain within the validation zone), which means that the dispersion variance of the normal scores within *Z* is 1.0, i.e.:

$$D^2(0 \mid Z) = \overline{\gamma}(Z,Z) = 1.0 \qquad (2)$$

where $\overline{\gamma}(Z,Z)$ is the average normal score variogram value within *Z*. The normal score variogram fit should be consistent with the above equality, which means that the variogram sill should be larger than one if the zone Z is not very large with respect to the variogram range, as it can be in the case of local grade control.

In practice, the variogram is often fitted first with a sill of one (Figure 1a, Step I.1.4). The value of $\overline{\gamma}(Z,Z)$ should then be computed. If the $\overline{\gamma}(Z,Z)$ value is within 5% of one, a simple rescaling of the variogram values is reasonable, otherwise a variogram model adjustment (sill and range) is suggested (Figure 4).

3.1.3 Gradual adjustment of simulated normal score average and variance.

If the source of the difference in mean (≠ 0.0) and/or variance (≠ 1.0) is unknown and there is reason to believe that the original N(0,1) distribution is correct, the simulated values may have to be adjusted. The following approach is a progressive correction that depends on the distance of the simulated node from the conditioning data.

First, a maximum possible adjustment at a given node $Sim_{tr\ max}(x)$ is defined by a simple standardization (mean = 0 and variance = 1):

$$Sim_{tr\max}(x) = (Sim(x) - Av_{Gsim}) / \sigma_{Gsim}$$

where $Sim(x)$ is the original simulated value at location x, Av_{Gsim} is the global average of all simulated values and σ_{Gsim} is the global standard deviation of all simulated values.

The actual adjustment $Sim_{tr}(x)$ is defined as follows:

$$\begin{aligned} &Sim_{tr}(x) = (Sim_{tr\max}(x) - Sim(x)) \cdot ratio(x) + Sim(x) \\ &ratio(x) = \sigma_{sim}(x) / \sigma_{\max sim} \end{aligned} \tag{3}$$

where $\sigma_{sim}(x)$ is the standard deviation of the simulated values at the selected node, and σ_{maxsim} is the maximum standard deviation of the simulated values from all nodes.

Note that for a node located on a conditioning data, $ratio(x)=0$ and $Sim_{tr}(x)=Sim(x)$, i.e., there is no correction. As the node gets further from the conditioning data, the value of $ratio(x)$ gradually increases from zero up to one, and the value of $Sim_{tr}(x)$ gradually varies from $Sim(x)$ to $Sim_{\text{tr}\ max}(x)$.

As shown in Figure 5, this adjustment results in a modification of both the average and the variance of the simulated values. Note that the adjustment does not result in an average and variance equal to 0 and 1 respectively, but there is a substantial improvement. Note also that the adjustment described in this section is a gradual affine correction that will not correct the shape of the distribution. If it is necessary to also correct the shape of the distribution (i.e., adjusting to a N(0,1) distribution), then a more sophisticated approach can be used (Xu, 1994).

3.2 VARIOGRAMS

The variograms of the simulated values can deviate from the modeled variograms. A deviation from the original model may adversely impact the simulation results, especially when the focus of the study is on variability of the mined blocks. The difference between the simulated and modeled continuities (variograms) may be caused by (1) poorly fitted variograms, (2) a modeler's decision to fit according to geological interpretation, and (3) unknown reason.

The first two sources of differences are counter-acted by data conditioning. Regardless of the original variogram model, continuities of the experimental data are to some extent imprinted on the simulated continuities, especially when there are lots of data as in mining. The impact of the data can be checked by comparing conditional and unconditional simulations. If deemed necessary, both variogram model range and sill may be adjusted to achieve the desired results. As shown in Figure 6 the adjustment to the variogram model results in improved, albeit not perfect, continuities of the simulated values.

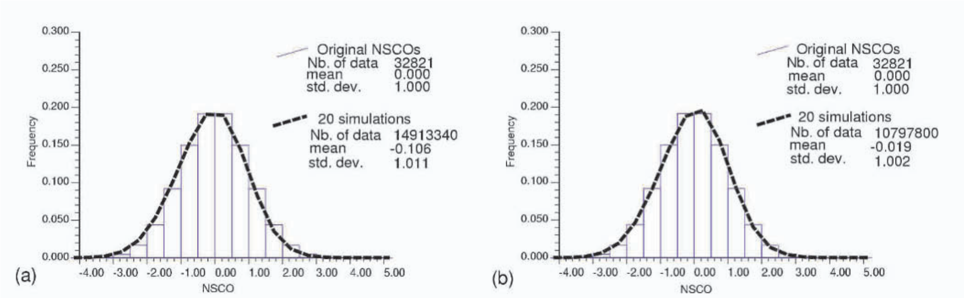


Figure 3. Comparison of simulated values with the data: (a) validation domain identical to simulation zone, (b) validation domain extending not further from the data than a search radius used for polygonal declustering

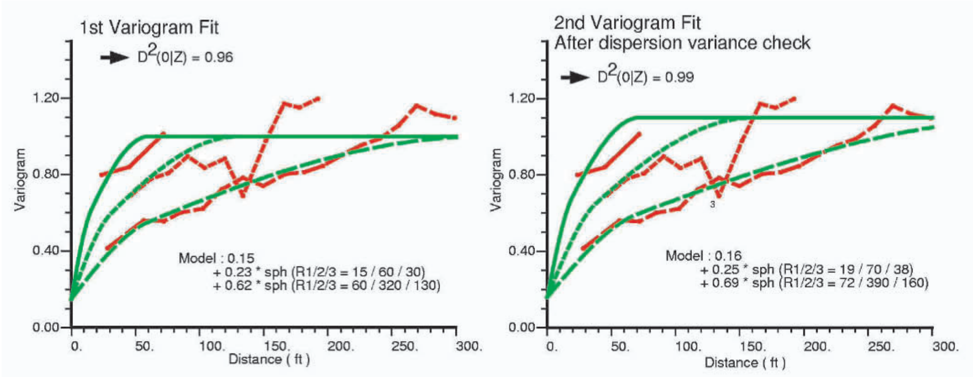


Figure 4. Example of variogram models before and after normal score variability check. a) Variogram model with total sill of 1.0 results in a dispersion variance within the validation zone of 0.96. b) Modified model with total sill of 1.10 results in a dispersion variance of 0.99.

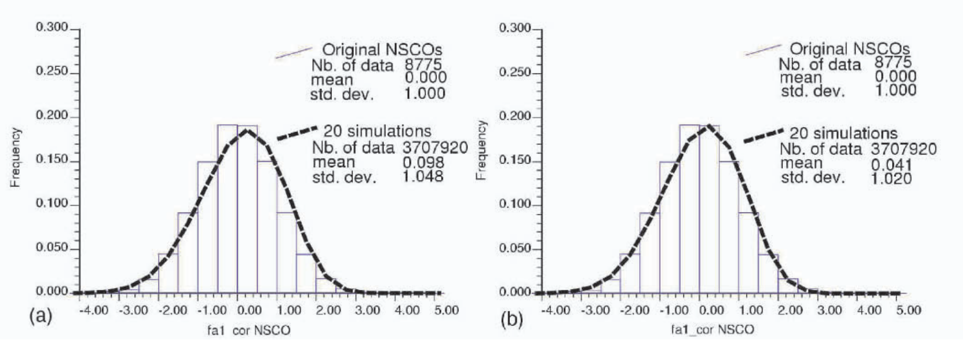


Figure 5. Comparison of simulated values with the original normal score data: (a) before the correction (b) after the correction of both average and variance

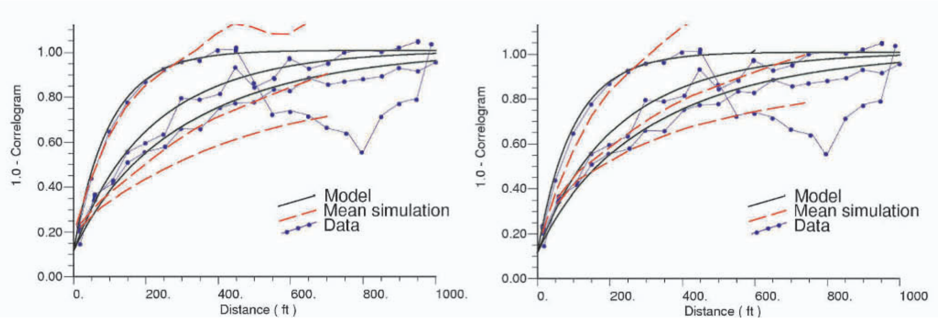


Figure 6. Variograms of simulated values (dashed curves) compared to the variogram models (solid curves) and to the experimental variograms (bulleted line) along different directions. Before correction for the sill and ranges (a), the simulated value variogram show more continuity than the experimental variograms. After correction (b), there is a better match between the two.

4 Check/Adjust simulated grades

All checks and sometimes the adjustments made in normal score space are necessary but not sufficient to ignore the checks on the simulated values after back-transformation (Step I.1.8 in Figure 1a). This is especially true because of a potential compounding effect of the corrections made. Although the writers are not aware of significant problems related to the series of corrections, their effect on the final simulated grades should be studied. Comparisons should be made with the original data within the validation envelope. Histograms, probability plots, scatterplots and visual checks of maps of simulated values are useful tools. Care should be given to ensure that the simulated mean grade in a geological domain is similar to the average estimated grade in that domain. If they are different, the simulated grades may have to be adjusted either by modifying some pre-simulation parameters, such as a trimming value, and re-simulating, or by a simple adjustment of the simulated values to the required average.

5 Bootstrap grades

Two main levels of uncertainty can be identified: geological (rock types) and grade uncertainty. Only the grade uncertainty is discussed in this section, but the same discussion applies to the geological uncertainty.

Current simulation practice often relies on the assumption that the distribution of in-situ grade values is known from the declustered grade histogram. The additional risk associated with an imperfect knowledge of the actual grade distribution should be addressed, resulting in better reproduction of the space of uncertainty.

Using a bootstrapping methodology, statistical fluctuations can be investigated by sampling from the original distribution (Steps I.2.1, I.2.2, and I.2.3 in Figure 1b). A typical procedure consists in creating a series of possible datasets by drawing randomly with replacement as many values, with the attached declustering weights, as there are in the original distribution. The fluctuations between the various datasets are then investigated.

When there are many sample values, such as in mining, the classical bootstrap approach results in datasets that are very similar to each other. This similarity would be perfectly correct if the sample values were uncorrelated, but this is not the case in a typical mining situation.

Spatial correlation can be addressed by drawing fewer values from the original distribution (Srivastava, pers. comm.). Indeed, the variance of the mean grade is:

$$Var1(\text{Mean}) = \frac{1}{N^2} \sum \sum C_{ij}$$

where C_{ij} is the covariance for the distance between sample i and j, and can be deduced from the variogram.

If P values are drawn randomly from the original dataset, the variance of the mean is:

$$Var2(Mean) = \frac{1}{P} Var(Data)$$

where Var(Data) is the variance of the original data set.

The required fluctuation for the mean is achieved if P is chosen such that Var2(Mean)=Var1(Mean), then:

$$P = \frac{Var(Data)}{Var1(Mean)}$$

Note that this formula could be refined to account for declustering weights.

Figure 7 illustrates the impact of bootstrap on the possible means of the original distribution. If no bootstrap is applied, the standard deviation of the mean is zero, i.e., the mean is fixed (Figure 7a). If the classical bootstrap is applied, the standard deviation of the means is 0.09 (Figure 7b). If the spatial bootstrap is applied, the standard deviation of the means increases to 0.036 (Figure 7c).

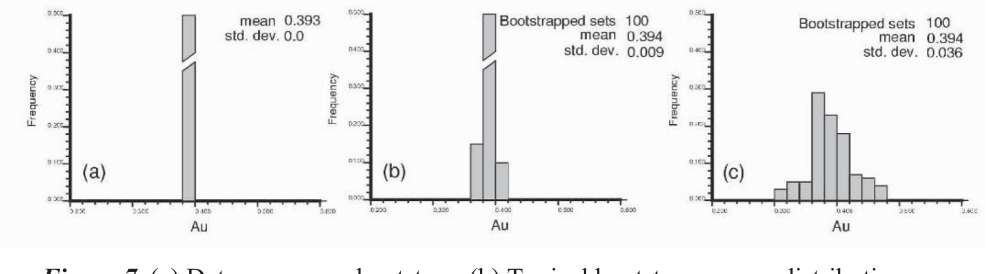


Figure 7. (a) Data mean - no bootstrap. (b) Typical bootstrap - mean distribution. (c) Spatial dependence bootstrap - mean distribution.

The bootstrapping may be done on data from all geological domains or on data from one domain at a time. If the former is used, the choice of optimistic (high average) and pessimistic (low average) distributions is more difficult, because the distributions from one or two domains may influence the results. The authors feel that bootstrapping per domain is a better solution. Under those circumstances, a pessimistic/optimistic distribution is truly pessimistic/optimistic in all domains. Of course, care should be given when choosing the bootstrapped distributions for simulating the grades. The distributions should not be overly pessimistic or optimistic. The choice of pessimistic/optimistic distributions can be limited to a specific area, or can be based on low/high metal content or NPV.

Prior to the final choice of the optimistic and pessimistic scenarios, it may be useful to have some insight on the potential impact of that choice on simulated values. Applying a cut-off grade on the bootstrapped distribution corrected for change of support may provide such insight.

Once a bootstrapped distribution is chosen, it is used first to generate a standard normal score transform (Figure 8a). The bootstrapped distribution and its transform are then used to convert the original grade values to normal score values (Figure 8b). The cumulative frequencies of the original sample grades are deduced from the bootstrapped distribution, then used to get the corresponding normal score values. Note that the resulting normal score values are not standard normal. For example, in the case of an optimistic bootstrapped distribution, the average of the normal score values of the original grade values is less than zero.

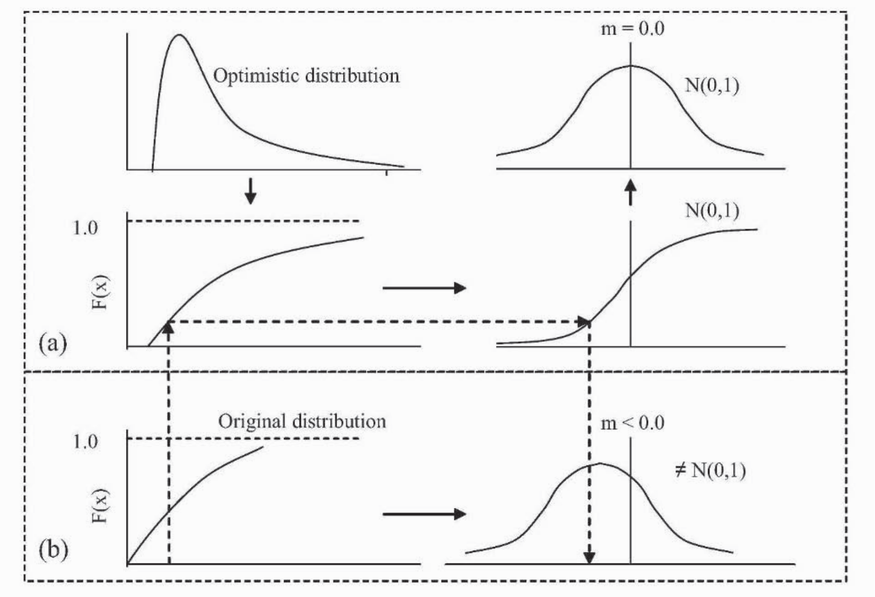


Figure 8. (a) Standard normal score transform based on a bootstrapped distribution. (b) Original grade distribution converted to normal scores using the standard normal score transform of the bootstrapped distribution.

6 Conclusions

A process for sequential Gaussian simulation is presented, which contains more steps than the usual normal score transformation, variogram modeling, simulation and back-transformation. A significant portion of the process may be used for other simulation methods, such as sequential indicator simulation. The authors believe that using similar processes in the mineral industry would avoid many costly mistakes.

Four of the most important aspects of the process are discussed in detail: trends, bootstrapping, checks, and adjustment of the simulated values.

Sequential Gaussian simulation often fails to correctly reproduce trends because of its strong stationarity requirement. A simple, albeit approximate, solution consists of adjusting for the trend after the simulation, via a gradual correction that depends on the distance to the conditioning data.

It is important that the simulation correctly reproduces the space of uncertainty. A modified bootstrap approach is presented to deal with the grade uncertainty. The modification is made to account for the spatial dependence between the samples. A similar approach can also be used to deal with the geological uncertainty.

To ensure high quality of the simulated values, a number of validation checks at different stages of the simulation are necessary. The checks start at the pre-simulation stage when experimental dispersion variances are compared against their theoretical values. Next, a series of checks followed by possible adjustments are done in the normal score space, and later similar checks and the adjustments are completed after back-transformation and trend addition. Most of the checks simply consist of comparing simulation and conditioning data statistics within a validation envelope. The potential adjustments are progressive, depending on the distance to the conditioning data.

References

Deutsch, C.V. and A.G. Journel, 1998, *GSLIB: Geostatistical Software Library and User's Guide*, Oxford University Press, New York, 380 pp.

Deutsch C.V., 2002, *Geostatistical Reservoir Modeling*, Oxford University Press, New York, 376 pp.

Goovaerts, P., 1997, *Geostatistics for Natural Resources Evaluation*, Oxford University Press, New York, 467 pp.

Isaaks, E.H., 1991. *Application of Monte Carlo methods to the analysis of spatially correlated data,* Unpublished PhD thesis, Stanford University.

Leuangthong O. and Deutsch, C.V., 2004. Transformation of Residuals to avoid Artifacts in Geostatistical Modelling with a Trend, *Mathematical Geology*, Vol 36, No 3, p. 287-305.

Nowak M., and Verly G., 2004. A Practical Process for Simulation, with Emphasis on Gaussian Simulation, Submitted to Orebody Modelling and Strategic Planning 2004 Symposium, Perth, Australia.

Xu, W., and Journel, A.G., 1994. Posterior identification of histograms conditional to local data. In *Stanford Center for Reservoir Forecasting Report* (SCRF) 7. Stanford Center for Reservoir Forecasting, School of Earth Sciences, Stanford, CA, USA

SPATIAL CHARACTERIZATION OF LIMESTONE AND MARL QUALITY IN A QUARRY FOR CEMENT MANUFACTURING

J. ALMEIDA[1], M. ROCHA[1] & A. TEIXEIRA[2]
[1]*CIGA, Centro de Investigação em Geociências Aplicadas, FCT/UNL, Monte de Caparica, 2829-516 Caparica, Portugal. ja@fct.unl.pt*
[2] *SECIL, Outão, Portugal*

Abstract. The aim of this study is to characterize the quality of the limestone and marl raw material exploited in a quarry for cement manufacturing by the SECIL Company (southern Portugal) based on the spatial distribution and variability of the chemical components (SiO_2, Al_2O_3, Fe_2O_3, CaO and MgO).
The first step of this study consists of the construction of sets of simulated images of these chemical components, using the Direct Sequential Simulation and Co-simulation algorithms. In the second step, the simulated images are combined on the quality indices LSF (lime saturation factor), SIM (silica modulus), ALM (alumina modulus) and CS (lime and silica ratio) in order to estimate local distribution laws of these indices. The local uncertainty and the probability of occurrence of extreme values are a tool of prime importance for the planning of temporal exploitation, regarding the proportioning optimisation mixture of raw materials coming from different quarry stopes.

1 Introduction

A set of parameters is currently used in cement manufacturing to characterize the quality of the raw material and to ensure the attendance of the quality of the produced cement. In Portugal, the SECIL Company uses four quality parameters (LSF – lime saturation factor; SIM - silica modulus; ALM – alumina modulus and CS – lime and silica ratio) and the magnesium grade (IPQ, 2001).
The LSF represents the relationship between the amount of calcium in the cement and the maximum amount theoretically possible for combining with other elements. It has a major influence in the manufacturing process and on the quality of the final product. An optimal LSF ranges between 1 and 1,02. It is calculated through the following relationship of grades, when expressed in weight percentage:

$$LSF = \frac{CaO}{1{,}8SiO_2 + 1{,}18Al_2O_3 + 0{,}65Fe_2O_3} \qquad (1)$$

The SIM is the second most important parameter to control the final product and it is calculated through the relationship between the grade of silica and the sum of the alumina and iron grades:

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 399-408.

$$SIM = \frac{SiO_2}{Al_2O_3 + Fe_2O_3} \tag{2}$$

A high SIM has the advantage of producing cement with high content of silicates, consequently with high mechanical resistance. Optimal values range between 2,4 and 2,6.

The ALM represents the relationship between the alumina and iron in the raw material; values should range between 1,5 and 1,7.

$$ALM = \frac{Al_2O_3}{Fe_2O_3} \tag{3}$$

Also the relationship between the calcium and the silica (CS) should be higher than 2:

$$CS = \frac{CaO}{SiO_2} \tag{4}$$

Finally, the magnesium grade (MgO) should be below than 5% in weight.

Raw materials exploited in marl and limestone quarries are combined amongst themselves to obtain optimal mixtures and if necessary with additives so that the final product presents quality parameters within adequate ranges.

Control of quality is done as soon as possible, starting from the quarry stopes. Samples collected from regular meshes of holes are chemically analysed by fluorescence of X rays on five chemical components: SiO_2, Al_2O_3, Fe_2O_3, CaO and MgO. Based on this regular but scarce information the main objective of this study is to provide images of the most probable values of these indices on each stope as well as the global and local uncertainty.

2 Methodology

In this work a stochastic simulation methodology is presented to characterize the quality of raw material within each stope according to the described quality parameters. The characterization of each chemical component and calculation of indices *a posteriori* instead of a direct characterization of the indices is preferable once the spatial variability in the quarry is mainly related to the deposition of each component and not on the quality indices themselves.

The main goal of the proposed methodology is to produce sets of images of five medium to highly correlated components, following the steps:

1. Exploratory data analysis of the chemical components in study: SiO_2, Al_2O_3, Fe_2O_3, CaO and MgO;
2. Application of Principal Component Analysis (PCA) and selection of the Principal Components (PC) that explain most of the initial variance;
3. Calculation of experimental variograms for the chemical components and PC selected and fitting of theoretical models;
4. Correlation analysis between chemical components and PC selected (calculation of correlation indexes);
5. Stochastic simulation of N_s images of the PC, using the Direct Sequential Simulation (DSS). These simulated images are used as secondary variables in the following steps of the proposed methodology.

6. Stochastic simulation of N_s images for each of the chemical components using the Direct Sequential Co-simulation. These images are conditioned to the experimental measurements (primary variable) and to the simulated images of PC (secondary variables).
7. Making use of the N_s simulated images of the chemical components, calculation of N_s correspondent images for each quality index (SIM, ALM, LSF and CS), following formulas (1) through (4) node by node. At each node location, the set of N_s values constitutes an estimation of the local histogram of each quality index giving the most probable value and the uncertainty.
8. Construction of probability maps showing areas where the quality indices (SIM, ALM, LSF, CS and the MgO grade) exhibit values in the optimal intervals;
9. Upscaling of the values defined at a small-scale block (step 8) to the stope boundary block size.
10. Construction of final indicator maps delimiting areas where the quality indices exhibit values in the optimal intervals.

2.1 BACKGROUND OF DIRECT SEQUENTIAL CO-SIMULATION WITH A SET OF SECONDARY VARIABLES

The DSS algorithm (Soares, 2001) was applied to produce simulated images of the *k* selected PC, respectively $Z_{PC1}(x)$, $Z_{PC2}(x)$,… $Z_{PCk}(x)$. Next step consists of the co-simulation of the chemical components conditioned to the previous simulated images of PC_1, PC_2, … PC_k as secondary images.

Direct Sequential Co-simulation with a set of secondary variables constitutes an extension of the initial algorithm proposed by Soares, 2001, and can be summarized as follows (Almeida *et al*, 2002):

1. Define a random path visiting each node of a regular grid of nodes.
2. At each node x_u, simulate the value $z^s(x_u)$ using the DSS algorithm:

   a) Identify the local mean and variance of $z(x)$ in x_u location, $z(x_u)^*$ and $\sigma^2_{sk}(x_u)$, using the simple co-located kriging estimator with a multiple set of secondary variables:

$$z(x_u) = \sum_{\alpha=1}^{n} \lambda_\alpha z(x_\alpha) + \sum_{i=1}^{k} \lambda_{PCi} z_{PCi}(x_u)$$

   Using the matrix formalism, the simple co-located kriging system with two secondary variables collocated in x_u and n neighbourhood samples is defined as follows (for sake of simplicity using two PC, PC_1 and PC_2):

$$\left[\begin{array}{cccc:cc} 1 & C_{12} & \cdots & C_{1n} & C_{1u}^{PC1} & C_{1u}^{PC2} \\ C_{21} & 1 & & C_{2n} & C_{2u}^{PC1} & C_{2u}^{PC2} \\ \vdots & & \ddots & \vdots & \vdots & \vdots \\ C_{n1} & C_{n2} & \cdots & 1 & C_{nu}^{PC1} & C_{nu}^{PC2} \\ \hdashline C_{u1}^{PC1} & C_{u2}^{PC1} & \cdots & C_{un}^{PC1} & 1 & C_{u}^{PC1PC2} \\ C_{u1}^{PC2} & C_{u2}^{PC2} & \cdots & C_{un}^{PC2} & C_{u}^{PC2PC1} & 1 \end{array}\right] \cdot \left[\begin{array}{c} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_n \\ \hdashline \lambda_{PC1} \\ \lambda_{PC2} \end{array}\right] = \left[\begin{array}{c} C_{1u} \\ C_{2u} \\ \vdots \\ C_{nu} \\ \hdashline C_{u}^{PC1} \\ C_{u}^{PC2} \end{array}\right]$$

Where:
$C_{\alpha\beta}$ - Covariance of the primary variable between samples at locations x_α and x_β
$C_{\alpha u}^{PC1}$ - Cross-covariance between primary variable at location x_α and PC_1 at location to estimate x_u
$C_{\alpha u}^{PC2}$ - Cross-covariance between primary variable at location x_α and PC_2 at location to estimate x_u
C_u^{PC1PC2} - Cross-covariance between secondary variables PC_1 and PC_2 at location to estimate x_u; equals zero.
λ_α - Weights of primary information
λ_{PC1} and λ_{PC2} - Weights of secondary variables
$C_{\alpha u}$ - Covariance of the primary variable between samples at locations x_α and location to estimate x_u
C_u^{PC1} - Cross-covariance between primary variable and secondary variable PC_1 at location to estimate x_u

and α=1…n; β=1…n (number of neighbouring samples of x_u).

b) Locally resample the histogram of $z(x_u)$, for instance using a normal score transform (φ) of the primary variable $z(x)$, and calculate $y(x_u)^*=\varphi(z(x_u)^*)$;
c) Draw a value p from a uniform distribution $U(0,1)$;
d) Generate a value y^s from $G(y(x_u)^*, \sigma^2_{sk}(x_u))$: $y^s= G^{-1}(y(x_u)^*, \sigma^2_{sk}(x_u),p)$;
e) Return the simulated value $z^s(x_u)= \varphi^{-1}(y^s)$ of the primary variable.

3. Loop until all nodes are simulated.

Assuming Markov-type approximation, the cross-covariance function can be calculated using the following relation in terms of covariance or correlograms (Almeida and Journel, 1994, Goovaerts, 1997):

$$C_{12}(h) \approx \frac{C_{12}(0)}{C_{11}(0)} C_{11}(h)$$

$$\rho_{12}(h) \approx \rho_{12}(0)\rho_{11}(h)$$

This approximation enables the inference of the primary variable is performed taking into account the spatial covariance of the primary variable and the correlation index between each secondary variable PC_1, PC_2, … and the primary variable (ρ_{PC1} and ρ_{PC2}):

$$C_{\alpha u}^{PC1} = \rho_{PC1}.C_{\alpha u}(h) \text{ and } C_{\alpha u}^{PC2} = \rho_{PC2}.C_{\alpha u}(h)$$

2.2 ZONATION OF RAW MATERIAL IN THE QUARRY STOPES

Direct Sequential Co-simulation produces simulated images of the five chemical components on a small-scale block. The set of N_s simulated values constitutes an estimation of the local cumulative distribution function for each variable within each node.

For each small block centred on location x_u quality parameters values were computed using formulas (1) through (4) obtaining *Ns* local values. For each parameter *p*, the N_s locally calculated values could be classified according to an indicator variable $I^p(x_u;N_s)$:

$$I^p(x_u;N_s)=\begin{cases}1 & \text{if } p(x_u,N_s)\in \text{optimal range for parameter } p\\ 0 & \text{otherwise}\end{cases}$$

where $p(x_u, N_s)$ is the value of parameter *p* in x_u calculated taking the N_s realization.

The average of the indicator values represent the proportion that small blocks belonging to the optimal range:

$$\frac{\sum_{N_s} I^p(x_u,N_s)}{N_s}=prob\quad p(x_u,N_s)$$

Two situations remain to solve: a) upscaling the small-scale block calculations to large-scale blocks at the same size of the stope (50m x 4m x 20m height) and; b) transform the probability values *prob* $p(x_u, N_s)$ into indicator values delimiting good and poor quality zones.

For a large block $v(u)$ constituted by N_v small blocks, it is estimated the following distribution law, which represents the proportion in volume of the large block $v(u)$ belonging to the optimal range for parameter *p*:

$$F^*(v_u,p)=\frac{1}{N_sN_v}\sum_{i=1}^{N_s}\sum_{j=1}^{N_v}I^p(x_j,i)$$

The final step consists on the calculation of probability thresholds p_c to transform the proportion maps $F^*(v(u), p)$ into binary maps:

$$I(v_u,p)=\begin{cases}1 & if\, F^*(v_u,p)>p_c\\ 0 & otherwise\end{cases}$$

The calculation of the threshold for parameter *p*, was based on the local and global probabilities of each large block to belong to each one of the categories (Soares, 1992, Almeida *et al*, 1993, Pereira *et al*, 1997).

3 Case study

The target area for exploitation is mainly constituted of grey and yellow limestone to marl (Kullberg *et al*, 2000). Orientation of the layers changes among N(80º to 90º)W in the most west area, N(65º to 70º)W in the central area and, approximately, N70ºW in the east area. Dip varies with more sloping at west (50º to 60º) than at east (≈ 45º).

A set of samples was extracted from a regular mesh of 131 vertical holes (Figure 1) for an area of 8 hectares. The total length of each hole is 20 meters (height of the steps of the exploitation), with a spacing of 10 meters in the N-S direction and 50 meters in the E-W direction. Each sample is a mixture of powder rock representing an average of grades for about 20 meters height, that validates the sample values in the characterization of three-dimensional blocks with the same vertical height.

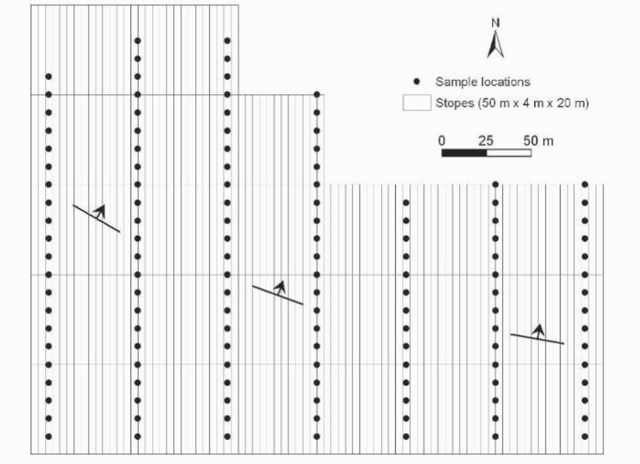


Figure 1. Spatial location of the samples in the studied area, orientation of the layers and design of stopes of 50 x 4 x 20 m^3 height.

Univariate statistics and correlation indexes were calculated for all initial data considered. The results are summarised in Tables 1 and 2.

It is observed a high positive correlation among SiO_2, Al_2O_3 and Fe_2O_3. The correlation between CaO/SiO_2, CaO/Al_2O_3 and CaO/Fe_2O_3 is also high, although negative. The MgO is not correlated with the remaining chemical components, meaning that its deposition is independent from the remaining components. This evidence is premonitory that two main PC will be necessary to synthesize all the initial information.

	Nº samples	Mean	Median	Variance	Skewness index
SiO_2		10,44	9,65	13,67	1,08
Al_2O_3		4,29	3,70	3,78	1,57
Fe_2O_3	131	2,11	1,90	0,49	1,48
CaO		41,60	41,47	13,11	-0,60
MgO		3,85	3,83	2,91	0,14

Table 1. Basic statistics of the initial dataset.

	SiO_2	Al_2O_3	Fe_2O_3	CaO	MgO
SiO_2	1,0000	0,9781	0,9468	-0,8631	-0,2403
Al_2O_3		1,0000	0,9586	-0,8161	-0,3103
Fe_2O_3			1,0000	-0,7653	-0,3566
CaO				1,0000	-0,2781
MgO					1,0000

Table 2. Correlation indexes between chemical components.

PCA algorithm was applied to synthesize the initial dataset to a reduced number of PC. Eigenvalues of the five axes and their explanation percentages are presented in Table 3. Figure 2 shows the graphical representation of the correlation indices between each initial variable and the PC and the projection of the 131 samples. According to Table 3 it is verified that first two PC transport more than 98% of the initial variance and that is enough to be used as secondary variables in the simulation of the images synthesizing the initial dataset.

Axis	Eigenvalues	Proportion of population variance (%)	Accumulated proportion (%)
1	3,711034	74,221	74,221
2	1,216618	24,332	98,553
3	0,052199	1,044	99,597
4	0,019663	0,393	99,990
5	0,000486	0,010	100,000

Table 3. Eigenvalues of each axis and proportion of population variance.

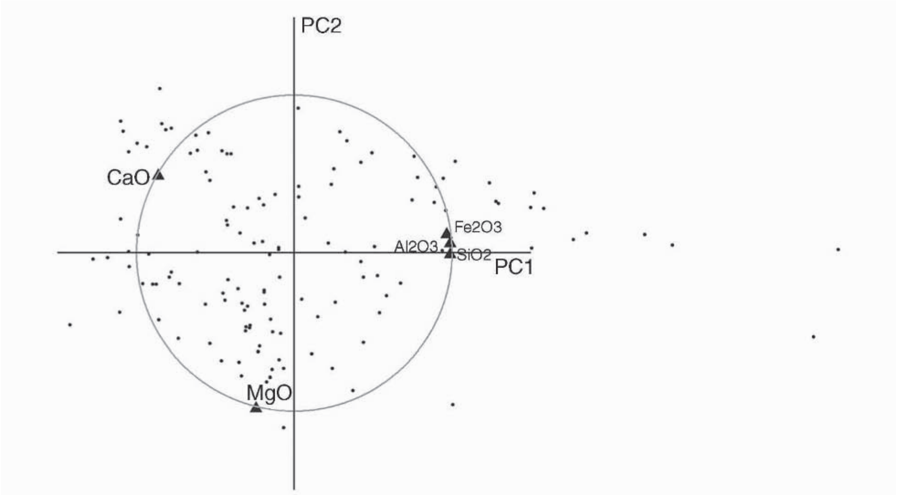


Figure 2. Graphical representation of the correlation indexes between initial variables and PC and sample projections on the two first PC.

	Model	C_1	a (N80˚W)	a (N10˚E)	Anisotropy
PC_1	Sph	0,74	200	30	6,67
PC_2		0,24	230	40	5,75
SiO_2	Sph	13,67	175	30	5,83
Al_2O_3		3,78	175	30	5,83
Fe_2O_3		0,49	250	35	3,33
CaO		13,11	125	20	6,25
MgO		2,91	200	60	5,83

Table 4. Models of variograms fitted to PC_1, PC_2 and chemical components.

Experimental variograms and fitting of theoretical models of spherical type (*Sph*) were made for the two selected PC and the five initial variables (see examples in Figure 3 and the parameters list in Table 4). Both PC_1 and PC_2 and all variables exhibit strongly anisotropic variograms (relationships between 3,33 and 6,67) where the main direction is related with the geological orientation of the layers.

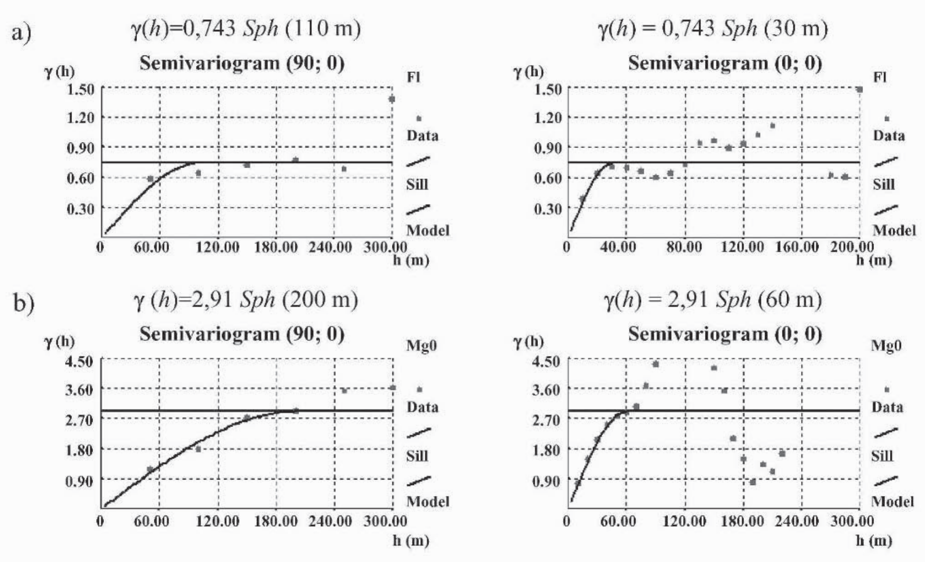


Figure 3. Experimental variograms and theoretical models fitted: a) PC_1; b) MgO.

The area in study was subdivided in a regular grid of 320 x 250 = 80000 small blocks with 1m by 1m length by 20 m height. Fifty images of PC_1 and PC_2 (to use as secondary variables) were simulated using DSS algorithm and fifty correspondent images of the initial variables were simulated using the proposed Direct Sequential Co-simulation algorithm conditioned to the PC_1 and PC_2 images.

The Figure 4 illustrates and example of a set of simulated images of PC_1, PC_2 and co-simulated images for each one of the initial variables. Each set of simulated images of the initial variables allows the calculation of a simulated image of the quality parameters as described in formulas (1) through (4). In order to validate the proposed method Table 5 shows the basic statistics for a set of simulated images.

	Number of blocks	Mean	Median	Variance	Skewness index
SiO_2	80000	10,45	9,59	14.20	1,29
Al_2O_3		4,29	3,81	4,22	1,30
Fe_2O_3		2,11	1,93	0,61	1,29
CaO		41,60	41,76	7,37	-0,71
MgO		3,86	3,82	3,30	0,90

Table 5. Basic statistics of one simulated set of images.

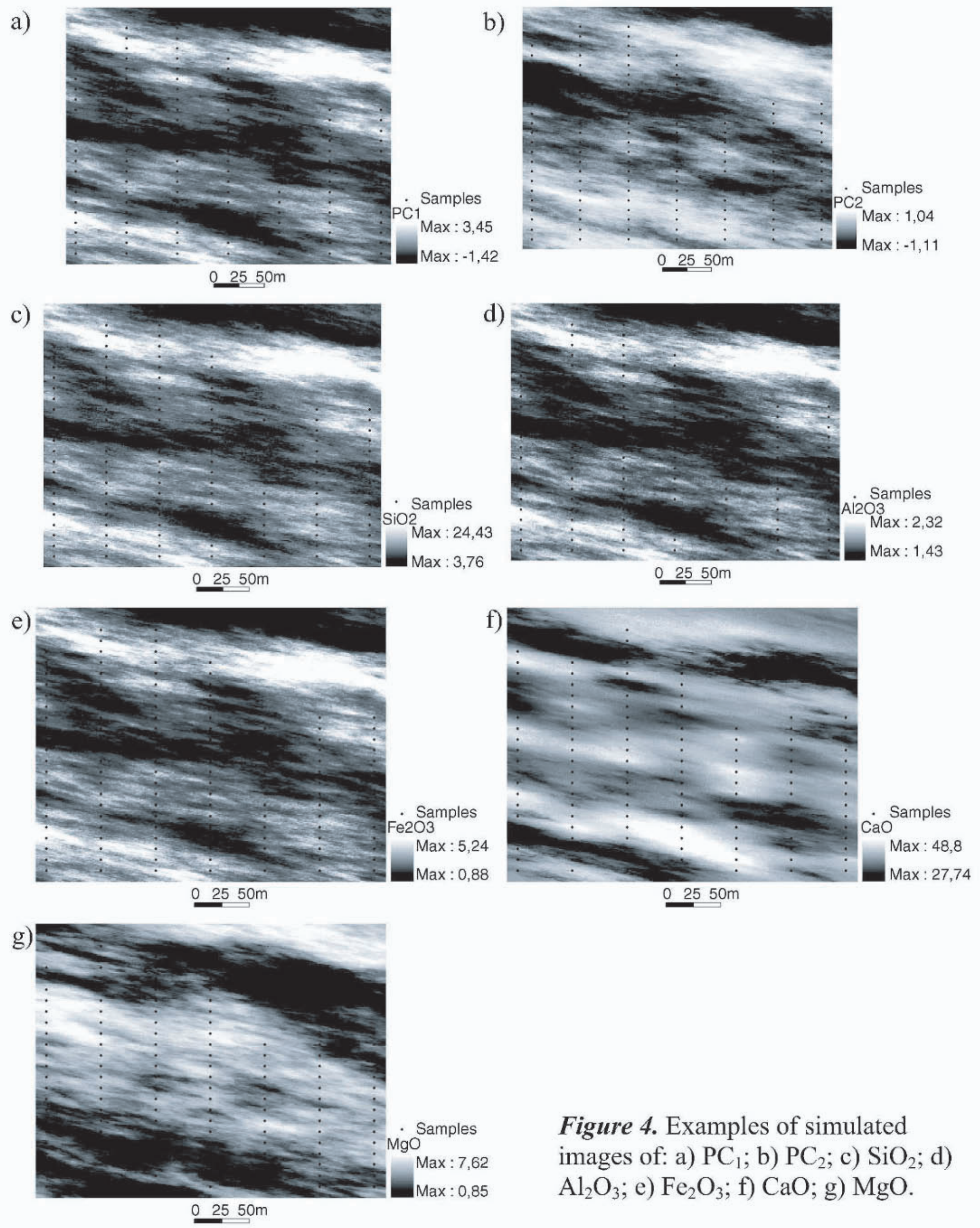


Figure 4. Examples of simulated images of: a) PC_1; b) PC_2; c) SiO_2; d) Al_2O_3; e) Fe_2O_3; f) CaO; g) MgO.

In order to upscaling all small blocks to a set of stope boundary blocks, 310 large blocks with 50 x 4 x 20 m^3 each were digitalized in the studied level of the quary (Figure 1). In the sequence of the proposed methodology local probabilities of each block to present parameters in the class of adequate quality were calculated and final probability maps were classified as indicator maps. For illustrative purposes, local probabilities and the limits of best areas are presented for the LSF and SIM parameters in Figure 5.

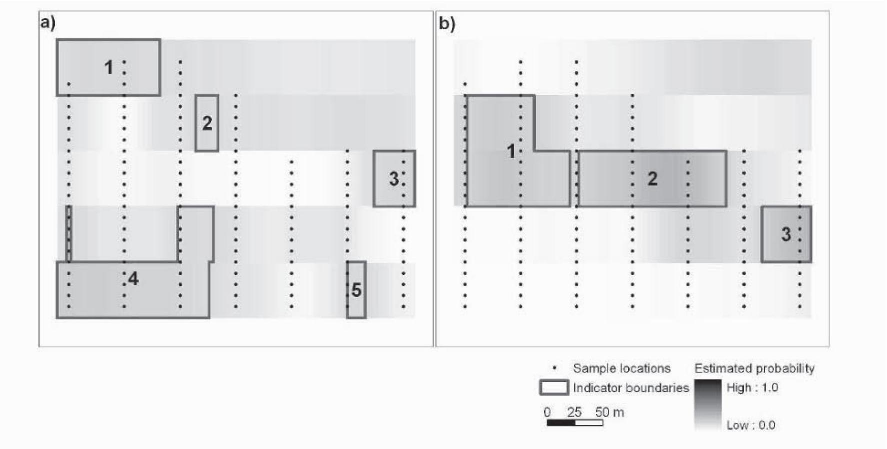


Figure 5. a) Probability of LSF $\in [0,66 ; 1,02]$ and identified areas; b) probability of SIM $\in [1,8 ; 3,0]$ and identified areas.

4 Conclusions

The presented case study shows a successful application of a multiple corregionalization simulation methodology, through the use of the main components of PCA as secondary variables and co-simulation of the main variables using these components as secondary variables. This methodology has the advantage of avoiding the problem of modelling multiple corregionalizations when a set of dependent variables is taken into account.

The final maps constitute an essential tool in the short-medium term planning of the exploration, allowing with a certain spatial resolution of the quality of the raw material exploited in each stope. The knowledge *a priori* of the most probable values and correspondent uncertainty of these chemical components and quality indexes in each exploitation step (local cumulative distribution functions), allow the optimal proportioning of raw materials, giving rise to a minimization of the costs namely addition of additives and stabilization of grades.

References

Almeida, A. and A. Journel 1994. Joint simulation of multiple variables with a Markov-type corregionalization model. Mathematical Geology 26(5), 565-588.

Almeida, J.; Soares, A. & Reynaud, R., 1993, Modelling the Shape of Several Marble Types in a Quarry, Proceedings of the XXIV International Symposium APCOM, Montreal, 3: 452-459.

Almeida, J.; Santos, E. Bio, A., 2002, Use of geostatistical methods to characterize population and recovery of Iberian hare in Portugal, Submitted to the Fourth European Conference on Geostatistics for Environmental Applications, Barcelona

Goovarets, P., 1997, Geostatistics for natural resources evaluation, Oxford University Press, New York, 483 p.

IPQ, 2001, Norma Portuguesa NP EN 197- 2001: Cimento: Composição, Especificações e Critérios de Conformidade para Cimentos Correntes, Instituto Português da Qualidade.

Kullberg M.C.; Kullberg J.C. & Terrinha P., 2000, Tectónica da Cadeia da Arrábida, in Tectónica das regiões de Sintra e Arrábida, Mem. Geociências, Museu Nac. Hist. Nat. Univ. Lisboa, 2, 35-84.

Pereria, M. J., Almeida, J, BritoG., Soares, A. and Zungailia, E. 1997, Stochastic Simulation of Sediment Quality for a Dredging Project, V. Pawlowsky-Glahn (ed.), Proceedings of IAMG'97, CIMNE, Barcelona, 2: 899-904.

Soares, A., 1992, Geostatistical estimation of multi-phase structures. Mathematical Geology, 24(2): 149-160.

Soares, A., 2001, Direct Sequential Simulation and Cosimulation. Mathematical Geology, 33(8), 911-926.

A NON-LINEAR GPR TOMOGRAPHIC INVERSION ALGORITHM BASED ON ITERATED COKRIGING AND CONDITIONAL SIMULATIONS

E. GLOAGUEN, D. MARCOTTE and M. CHOUTEAU
Department C.G.M., C.P. 6070 succursale centre-ville, Montréal, Québec, Canada

Abstract. A new constrained velocity tomography algorithm based on ray approximation is presented. This algorithm is based on slowness covariance modeling using experimental travel time covariance. The computed covariances, the measured travel times and additional slowness values allow cokriging and conditional simulation. Among several realizations, the one that minimized the L1 norm is chosen as the best velocity field. In the proposed method the raypaths must be known. Starting with a homogeneous velocity field, an iterated solution is computed updating the raypaths applying Snell-Descartes' law on the best velocity field after each iteration. First, the advantage of an iterated solution is presented. Then, the proposed approach is compared to a classical LSQR algorithm using a synthetic model and real data collected for geotechnical evaluation in a karstic area. The tomographies on synthetic models show that geostatistical methods provide comparable to or better results than LSQR. For both methods, additional velocity constraints reduce uncertainty and improve spatial resolution of the inverted velocity field. Also, the simulation on synthetic models increases the spatial resolution compared to LSQR. The real data analysis shows that the proposed method gives very consistent results with respect to the drilling log information.

1 Introduction

Ground Penetrating Radar (GPR) is a non-destructive geophysical technique which uses radio waves (10 to 2000 MHz) to investigate electrical properties of the ground. A popular method of GPR data acquisition is cross-hole tomography. The transmitter, located in one hole, emits an electromagnetic wave impulse. The travel time to a receiver located in a coplanar hole is recorded. The goal is to determine the spatial distribution of slowness from the different travel times, a fundamentally non-linear problem. Common approaches discretize the plane between the holes in a series of cells in each of which the slowness is considered constant (Holliger et al., 2001). Commonly used tomographic algorithms (LSQR (Paige and Saunders, 1982)) use the ray approximation for wave propagation. These algorithms require the user to specify critical parameters often obtained only by trial and error. We

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 409-418.

propose an iterative method based on a stochastic model for the cell slowness. The first step consists of identifying the slowness stationary covariance structure from the non-stationary travel times. This is accomplished by using the ray approximation for wave propagation. The covariances and cross-covariances are linearly related through a geometric matrix describing the paths. The second step consists of simple cokriging and conditional simulation of the slowness field from the arrival times and the ray approximation. Travel times modeling is performed using all the simulated velocity field. The simulation that produces the travel time vector that minimizes the L1 norm compared to the measured travel times is used to compute new propagation paths applying Snell-Descartes' law. The system is updated and a new solution is obtained. Usually, after a few iterations the solution obtained is stable. Moreover, a substantial reduction in travel time Mean Square Error is observed with these final simulations compared to the classical cokriging solution or to alternative inversion algorithms.

It is possible to have velocity information along the holes, for example, using borehole reflection surveys. LSQR and geostatistical methods allow including these additional data. This allows a dramatic increase in the spatial resolution and also decreases the uncertainty on velocity estimates. First, the GPR technique is presented and a classical tomography method is briefly described. Then, the theory of the proposed method is presented. The proposed method and classical tomography are compared using a synthetic stochastic model. Finally, LSQR and the geostatistical method are used to image a karstic zone in a geotechnical study.

2 Ray based tomography

An easy way to approximate a wave path in propagation mode is to use the ray. A ray is defined as the curve that connects a transmitter to a receiver, and lies perpendicular to the wave front (Berryman, 2000). For ElectroMagnetic propagation, the ray geometry depends on the electric property contrasts, and, thus, on the velocity contrasts as described by Snell-Descarte's law.

In ray-based tomography, the field is discretized as a series of cells. For each transmitter-receiver pair, the length of each segment of ray path that crosses a cell is computed. All the segment lengths are organized in a (sparse) matrix L, called the parameter matrix, which describes the geometry of the rays. L is of size nt observed times by np cells (of constant slowness). Equation 1 represents the linear relation between travel time vector t and the slowness vector s.

$$Ls = t \tag{1}$$

This equation represents the forward modeling of the travel time. The slowness field "the unknown" must be estimated by inversion of Equation 1.

2.1 REGULARIZATION AND EQUALITY CONSTRAINTS IN CLASSICAL INVERSION

Generally, in Equation 1, L cannot be inverted directly. In most cases, the problem is ill-posed. The linear system is modified to include a regularization term (Menke, 1989).

$$\begin{bmatrix} L \\ kD \end{bmatrix} \begin{bmatrix} s \end{bmatrix} = \begin{bmatrix} t \\ 0 \end{bmatrix} \tag{2}$$

where k is a scalar and D is typically the discrete first derivative (flatness) of the slowness field. D can also be taken as the identity matrix (smoothness). The solution can be both smoothed and flattened by taking a weighted sum of the identity matrix and the derivative matrix.

When slowness values are known within the field that is to be inverted, it is suitable to force solution to fit the known values. The implementation of such equality constraints is easy in linear systems (Menke, 1989). Equation 1 is modified to take into account the velocity constraints:

$$\begin{bmatrix} L \\ M \end{bmatrix} \begin{bmatrix} s \end{bmatrix} = \begin{bmatrix} t \\ sc \end{bmatrix} \tag{3}$$

where M is a matrix of size sc x np, sc is the vector of known cell values. In each row, M is equal to one in the column corresponding to a known value and zero elsewhere.

In this study, the LSQR algorithm (Paige and Saunders, 1982), a classical tomography algorithm is used. This is a conjugate gradient type algorithm with Golub-Kahan bidiagonalisation (Berryman, 2000). The algorithm converges quickly and is particularly effective for sparse matrices. However, the convergence criteria must be carefully chosen to avoid the algorithm iterating on noise. Here, the correlation from one iteration to the next, and the derivative of the sum of the residuals were used as convergence criteria. Flatness and smoothness regularizations were combined.

2.2 PROPOSED METHOD

The stochastic approach for linear system inversion was first presented in Franklin (1970). Being linearly related, slowness and travel time covariance matrices are also linearly related. The linear relation between slowness and travel time covariances is:

$$cov(t,t) = Lcov(s,s)L^T + C_0 \tag{4}$$

where $cov(t,t)$ is the nt x nt travel time covariance matrix, $cov(s,s)$ is the np x np slowness parameter covariance matrix and C_0 is the travel time nugget effect.

The covariance matrix for the slowness is specified by choosing the model function and its parameters (nugget, sill, and range). Once the model function is selected, the slowness covariance parameters are estimated by an iterative search in the low-dimension covariance parameter space.

2.3 COKRIGING

Cokriging (Chilès and Delfiner, 1999) is a mathematical interpolation and extrapolation tool that uses the spatial correlation between a secondary variable (here, the measured travel times) and a primary variable (here, the slowness) to improve estimation of the primary variable at unsampled locations. When an acceptable slowness covariance model is obtained, the slowness field is cokriged using the arrival times and any available slowness data. It is also easy to go further and to impose slowness gradients or even any kind of linear constraint to the solution. The simple dual cokriging weights Γ are given by:

$$\Gamma = \begin{bmatrix} cov(t,t) & cov(t,sc) \\ cov(sc,t) & cov(sc,sc) \end{bmatrix}^{-1} \begin{bmatrix} t \\ sc \end{bmatrix} \tag{5}$$

where *cov* signifies covariance, *sc* are the known slowness cells and *s* are the slowness cells that are to be inverted.

The cokriging estimator Z_g^* for the slowness is given by:

$$Z_g^* = \Gamma^T \begin{bmatrix} cov(t,s) \\ cov(sc,s) \end{bmatrix} \tag{6}$$

2.4 SIMULATION

By construction, cokriging gives a smooth estimate of the slowness field. It may be desirable and informative to obtain various reasonable solutions showing the kind of variability that can be expected from the slowness covariance model adopted. This is obtained by using geostatistical simulation algorithms rather than cokriging. There exist many efficient simulation algorithms (Chilès and Delfiner, 1999). The Fast Fourier Transform Moving Average simulation (FFT-MA) is a fast simulation algorithm for generating regular grid non conditional Gaussian stationary processes (Le Ravalec et al., 2000). Conditioning of FFT-MA simulation is performed by cokriging, using the same weights as in Equation 5. For each tomography, several simulations are computed. For each simulation, travel times are computed using Equation 1. The best simulation is defined as the one that minimizes the sum of the absolute difference between the computed and the measured times (L1 norm). Contrary to LSQR, both cokriging and conditional simulation retrieve exactly the slowness data.

3 Results

3.1 ADVANTAGES OF CURVED RAYS

Because the true velocity field is not known, the first iteration is performed using straight ray approximation. Of course, the straigth ray is not a satisfying approximation. Figure 1 and Figure 2 show the influence of low and high velocity anomalies along the raypath, respectively. Intuitively, a low velocity anomaly appears smaller in the straight ray reconstructed images because the ray convergence toward the high velocity zone. Conversely, a high velocity anomaly appears larger in the reconstructed images than in reality. A well known technique is to update the raypath after each iteration taking into account the velocity cell constrasts (Berryman, 2000).

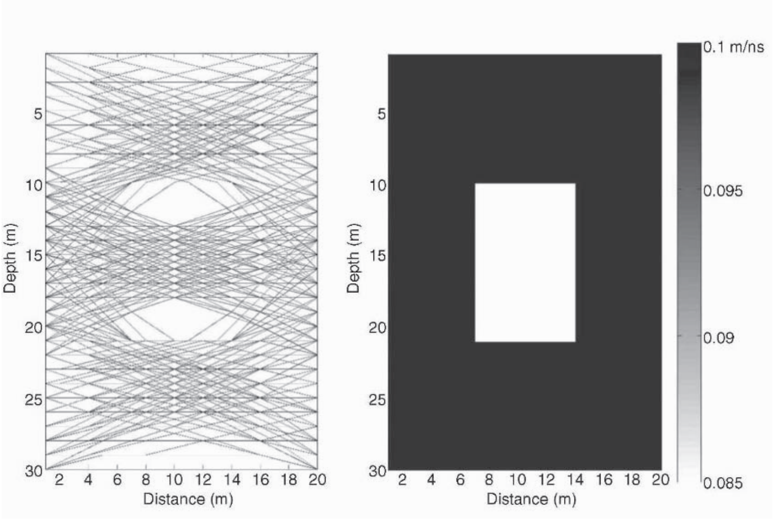


Figure 1. Left: raypaths. Right: Low velocity anomaly.

Figure 3 shows 6 iterations of the proposed method on a synthetic model. The model consists of a rectangular anomaly (represented by white dashed line) of 0.125 m/ns in a medium of 0.1 m/ns. The optimized covariance model is an isotropic spherical model. The range is 6 m, the slowness sill is 3 (ns/m) ^{2} and the travel time nugget effect is 1 ns^2. The covariance model stays the same for all the iterations. For each iteration the best of 100 simulations is chosen. Figure 3 shows that curved ray tomography allows an increase in the spatial resolution and reduces numerical artifacts. After only one curved ray iteration, the reconstructed anomaly is well recovered. At the fourth iteration, there remains only few artifacts. Figure 4 shows the L1 norm of 20 iterations. This figure illustrates that after the fourth iteration there is no improvement in the reconstructed image.

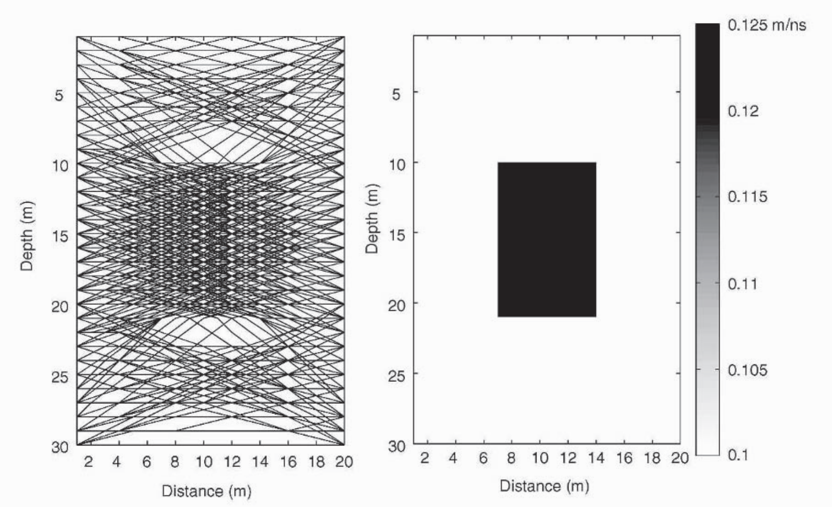


Figure 2. Left: raypaths. Right: High velocity anomaly.

3.2 TOMOGRAPHY ON SYNTHETIC DATA

In this section, the results of constrained LSQR and simulation tomographies on a synthetic velocity model are presented. The synthetic velocity model is presented in Figure 5a. Transmitter and receiver positions are also plotted in Figure 5a. A curved ray tracing algorithm based on graph theory was used to compute synthetic travel times (Berrymann, 1991; Moser, 1991). The modeled slowness covariance function is Gaussian with ranges 6 m along the horizontal axis and 3 m along the depth axis. The travel time nugget effect is 1 ns^2 and the slowness sill is 0.2 (ns/m)^2. The velocity in every cell intersected by the two boreholes is fixed as a constraint. Velocity constraints are implemented as presented in Equations 3, 5 and 6.

Figures 5b and 5c present the LSQR and the best simulation tomographies. For both methods, the main features of the velocity field are recovered. But, iterative simulation allows an improvement in the spatial resolution. The correlations between the velocity model and tomography images are 0.73 and 0.89 for the constrained LSQR and the iterated simulation, respectively.

4 Geotechnical evaluation in a karstic area

Borehole GPR measurements were performed to complement the site characterization of a planned expansion of a cement plant, including a mill and a reclaim facility adjacent to existing buildings. The whole site is located in a karstic environment. The overburden is an irregular residual clay layer overlying a limestone bedrock. Sixteen holes were visited during the survey (Figure 6). A RAMAC system with 100 MHz borehole antennas was used for the survey. Single-hole reflection mea-

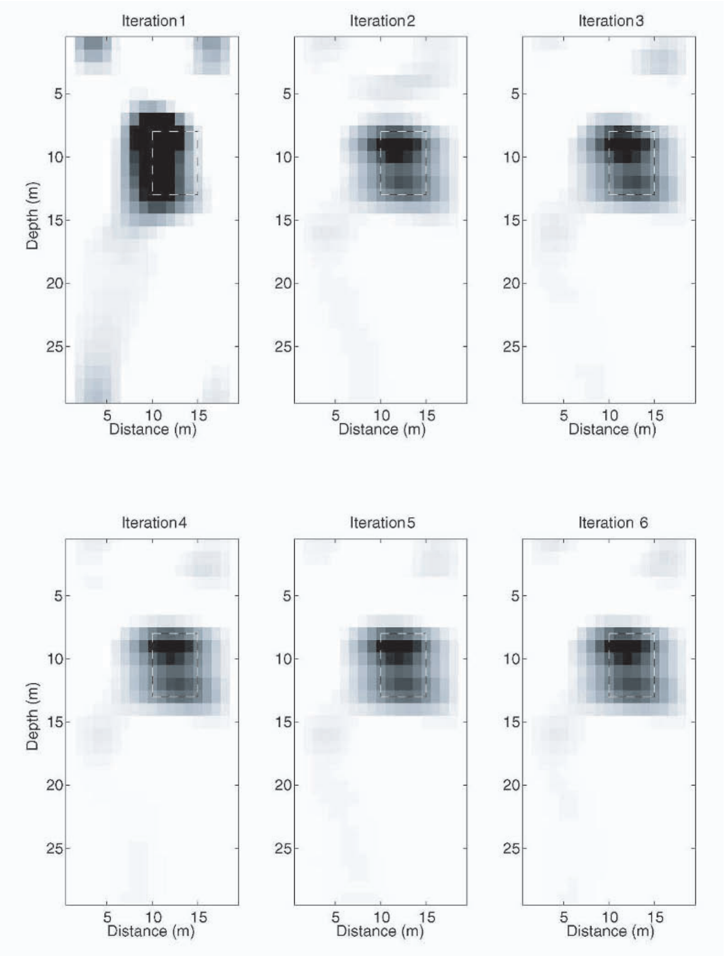


Figure 3. Curved ray iteration for the proposed method on block synthetic model

surements were performed in each hole and nineteen tomographic panels were acquired. In this article, four holes have been used to perform velocity tomography (AR13, AR08, AR12 and AR18 in Figure 6). Because they are nearly coplanar, they were included in the same 2D tomography. Slowness constraints were obtained by inversion of single-hole radar profiles (Giroux et al., 2004). Figure 7a shows the result of constrained LSQR tomography and the stratigraphy obtained from drilling logs. Figure 7b shows the constrained simulation that minimizes the L1 norm. The modeled slowness covariance function is Gaussian. The ranges are 15 m along the horizontal axis and 6 m along the depth axis. The travel time nugget effect is 10^{-6} ns^2 and the slowness sill is 1.5 x 10^{-6} (ns/m)^2. The stratigraphy is also shown. It is clear, that the proposed method offers a better match with drilling log information. Also, the conditional simulation provides an image easier

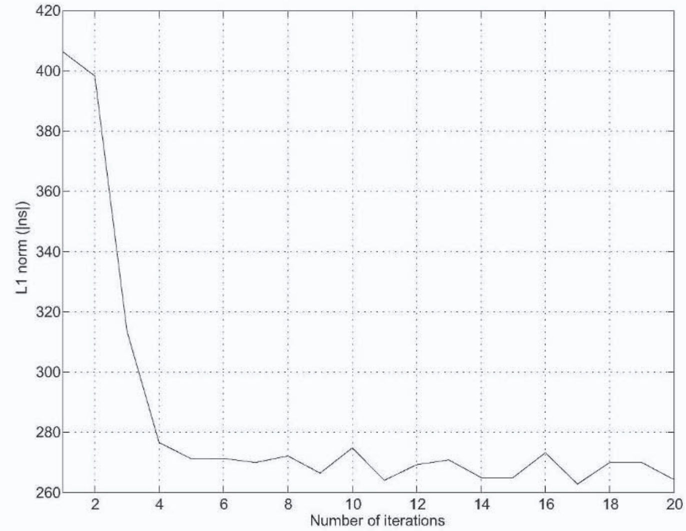


Figure 4. Objective function for block synthetic model

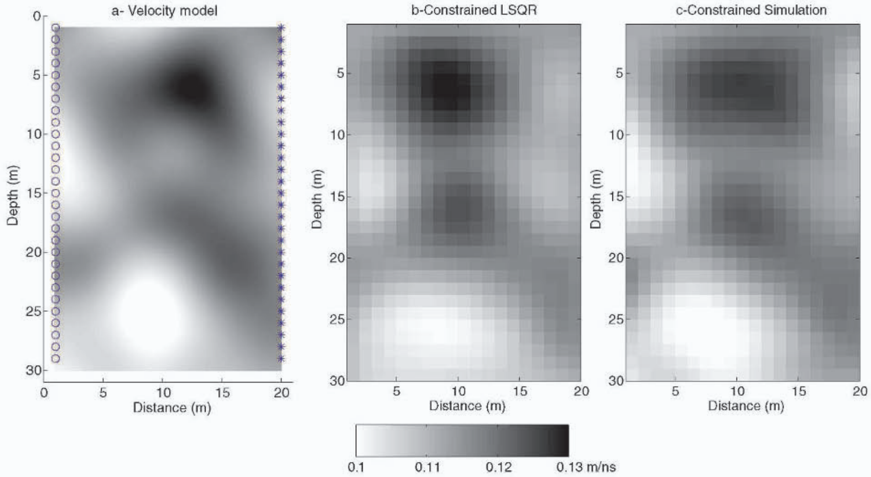


Figure 5. a- velocity model (o: transmitter, ⋆: receiver). b- constrained LSQR. c- constrained simulation

to interpret geologically. As expected, the velocity of the clay is lower (about 0.07 m/ns) than the velocity of the limestone (about 0.12 m/ns). These values compare well with the theoretical ones (Dubois, 1995; Feschner et al., 1998). Moreover, strong artifacts are generated by LSQR that render the interpretation difficult.

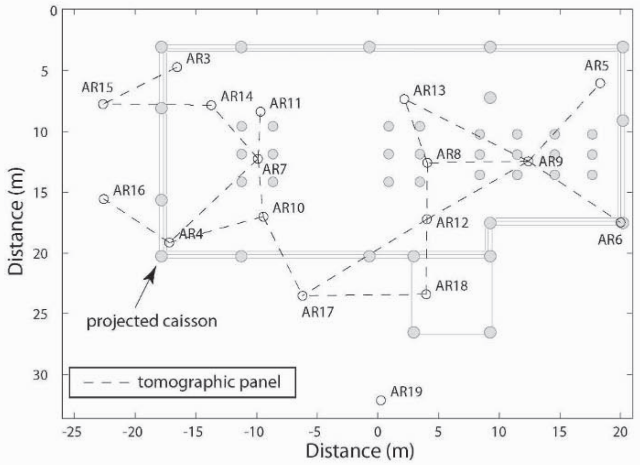


Figure 6. Borehole locations in the survey site

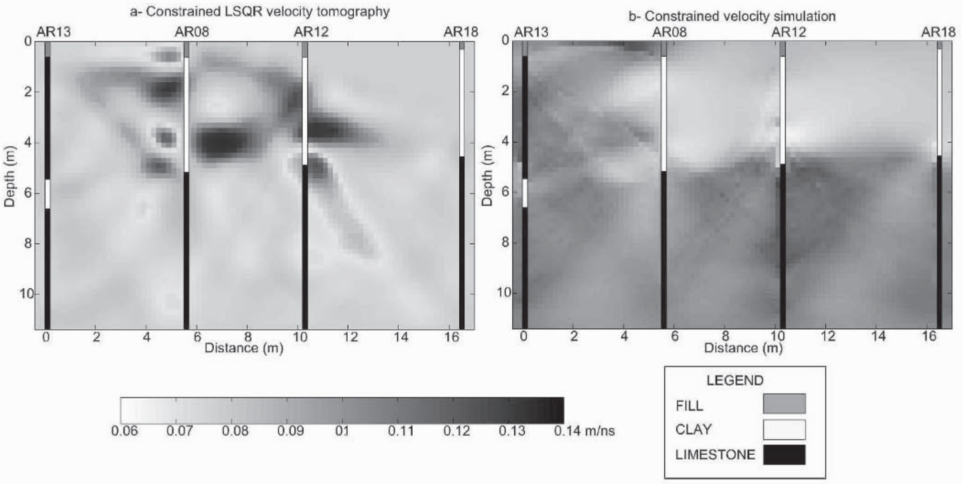


Figure 7. a-Constrained LSQR b-Conditional simulation

5 Conclusions

It has been demonstrated that the geostatistical tomography gives similar to or better results than LSQR. Moreover, the proposed method allows to take into account the non linearity of raypaths. During the real data analysis, simulation allows finding a velocity field in excellent agreement with the drilling log information.

Acknowledgements

We would like to thank Dr. Bernard Giroux for providing his ray path modeling code. We also thank Maria Annecchione for editing this manuscript.

References

Berryman, J.G., *Analysis of approximate inverses tomography*, Optimization and Engineering, vol. 1, 2000, p. 437-473.

Chilès, J. P. and Delfiner, P., *Geostatistics: Modeling spatial uncertainty*, Wiley Series on Probability and Statistics, 1999.

Dubois, J-C., *Borehole radar experiments in limestone: analysis and data processing*, *First Break*, vol.13 no. 2, 1995, p. 57-67.

Fechner, T., Pippig, U., Richter, T., Corin, T., Halleux, L. and Westermann, R., *Borehole radar surveys for limestone investigation*, *in Proc. GPR1998*, Lawrence, Kansas, USA, 1998, p. 127-132.

Franklin, J. N., *Well-posed stochastic extensions of ill-posed linear problems*, J. Math. Anal. Apll., vol. 31, no. 1, 1970, p. 682-716.

Giroux, B., Gloaguen, E. and Chouteau, M., *Geotechnical application of borehole GPR - a case history*, In Proc. GPR2004, Delft, The Netherland, 2004, p. 249-252.

Holliger, K., Musil, M. and Maurer, H.R., *Ray-based amplitude tomography for crosshole georadar data*, Journal of Applied Geophysics, vol. 47, 2001, p. 285-298.

Le Ravalec, M., Noetinger, B. and Hu, L. Y., *The FFT Moving Average generator: an efficient numerical method for generating conditioning gaussian simulation*, Mathematical Geology, vol. 32, 2000, p. 701-723.

Menke, W., *Geophysical data analysis*, Academic Press, 1989.

Moser, T. J., *Shortest path calculation of seismic rays*, Geophysics, vol. 56, 1991, p. 59-67.

Paige, C.C. and Saunders, M.A., *LSQR: an algorithm for sparse linear equations and least-squares*, ACM Trans. Math. Soft., vol. 8, no. 1, 1982, p. 43-71.

APPLICATION OF CONDITIONAL SIMULATION TO QUANTIFY UNCERTAINTY AND TO CLASSIFY A DIAMOND DEFLATION DEPOSIT

SEAN DUGGAN [1] AND ROUSSOS DIMITRAKOPOULOS [2]

[1] *MRM-Placers, De Beers Group Services, Aon House, 117 Hertzog Blvd, Cape Town, South Africa.*

[2] *WH Bryan Mining and Geology Research Centre, University of Queensland, Brisbane, Australia*

Abstract. Since the early 1900's diamonds have been known to occur in aeolian placers in south western Namibia. At Namdeb's Elizabeth Bay Mine diamonds are extracted from the fine to coarse grit layers in a sequence of stratigraphic horizons formed during periods of vigorous wind action. Significant capital expenditure is required to extend the life of mine at Elizabeth Bay and, as this is an inherently high-risk deposit, a sound understanding of the risks associated with the resource estimates is required. Various methods were evaluated to quantify the uncertainty of the thickness estimates and to facilitate classification according to the SAMREC guidelines. The thickness of the resource has a significant impact on the mining method as well as volume calculations. This investigation involves the use of conditional simulation of thickness to derive a method for classifying the resource. The simulations were used to construct block conditional distribution functions and evaluate a number of uncertainty measures, including conditional variance, conditional coefficient of variation, interquartile range and probability interval. A method employing conditional simulation to assess the efficiency of sample spacing is briefly presented. The approaches using coefficient of variation calculations provide promising results that enable classification of uncertainty related to the thickness of the Elizabeth Bay resource.

1. Introduction

Diamonds were discovered in a contemporary aeolian placer in the vicinity of Lüderitz in 1908. Recent mining, from 1990 to the present, has focussed on the Quaternary to Recent aeolian placers at Elizabeth Bay, c. 40km south of Lüderitz where the economic horizons comprise siliceous grits to small pebble size beds (mostly 2 - 8mm clast size). In order to continue mining into the future an additional capital expenditure is required to extend the life of mine and in addition to mining and treatment difficulties there is a risk associated with the uncertainty of the estimate of grade, average diamond size and resource thickness. This study addresses uncertainty related to the thickness of the economic part of the deposit.

Good spatial characteristics make the use of geostatistical estimation methods possible for most variables including resource thickness. The diamonds are found in a sequence of fine to coarse grit horizons formed during periods of vigorous wind action and, in parts, fluvial reworking. The mineralised component of the deposit comprises a thin upper deflation grit known locally as Grey Beds, overlying a thicker sequence called

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 419-428.

Brown Beds. In the north-east of the deposit the Brown Beds are underlain by the Red Beds (or Fiskus Sandstone), but because of limited representative sampling Red Bed data was excluded from this study.

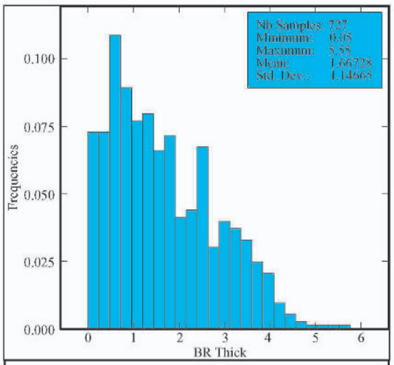
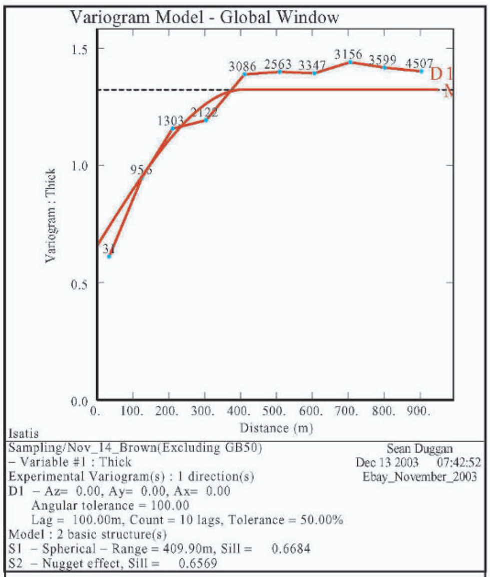


Figure 1: Histogram of Brown Beds thickness; sample data

Although Namibia does not have a specific code for classification of resources the guidelines outlined in the South African Code for Reporting of Mineral Resources and Mineral Reserves (SAMREC) are accepted by most mining companies. However, this code, like other international codes, provides only broad guidelines and is in no way quantitative.

2. Resource Thickness

The thickness of the ore body determines the mining method used (and therefore the carats recovered) and is used to obtain a local estimate of volume. Ordinary kriging was used to estimate grade, average diamond size and resource thickness in the three main stratigraphic horizons into 100m x 100m blocks.



```
Sampling/Nov_14_Brown(Excluding GB50)                Sean Duggan
– Variable #1 : Thick                          Dec 13 2003   07:42:52
Experimental Variogram(s) : 1 direction(s)       Ebay_November_2003
D1  – Az=  0.00, Ay=  0.00, Ax=  0.00
     Angular tolerance = 100.00
     Lag =  100.00m, Count = 10 lags, Tolerance = 50.00%
Model : 2 basic structure(s)
S1  – Spherical – Range = 409.90m, Sill =    0.6684
S2  – Nugget effect, Sill =    0.6569
```

Figure 2 : Brown Bed thickness variography

The average thickness of the Grey Beds at Elizabeth Bay is 0.60m and the underlying Brown Beds are, on average, 1.67m thick. The histogram derived from sample data for the Brown Bed horizon is shown in Figure 1.

The Brown Bed thickness semi-variogram (Figure 2) shows an isotropic, Spherical model with a range of 410m and a nugget effect of about 50%. A similar model was obtained for the Grey Beds with a double structure with ranges of 125m and 500m. Ordinary kriging was used to estimate the average thickness of both horizons using 100m x 100m blocks.

Figure 3 illustrates the estimated resource thickness for the Grey Beds and the estimation standard deviation (Figure 3, right). The latter reflects the sample density (black dots), as expected.

3. Method

A summary of the method used to find a measure of uncertainty is outlined in Figure 4. The technique requires successful conditional simulation realizations of the variable under study on a dense grid of nodes (data support). Reproduction of the data histogram and variogram model are validated for each realization. The change of support of the realizations into the required block size is performed or alternatively, simulated directly on blocks (e.g., Godoy, 2003). The second part of the method is to derive the conditional distribution functions (cdf's) from the set of simulated resource models, describing the uncertainty about the unknown thickness values for each block. The third step involves computing the uncertainty measure(s) from the cdf of each block and finally classifying the blocks into a specific category of resource by selecting thresholds using the set of available, simulated ore body models.

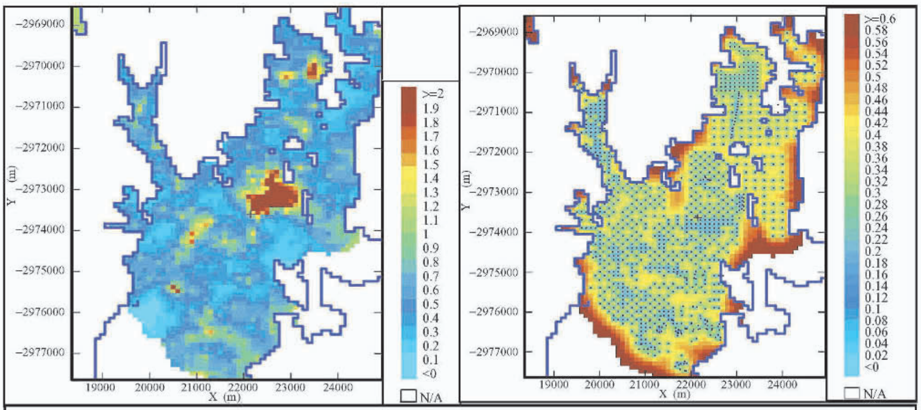


Figure 3 : Ordinary kriged estimate (left) left and estimation standard deviation (right)

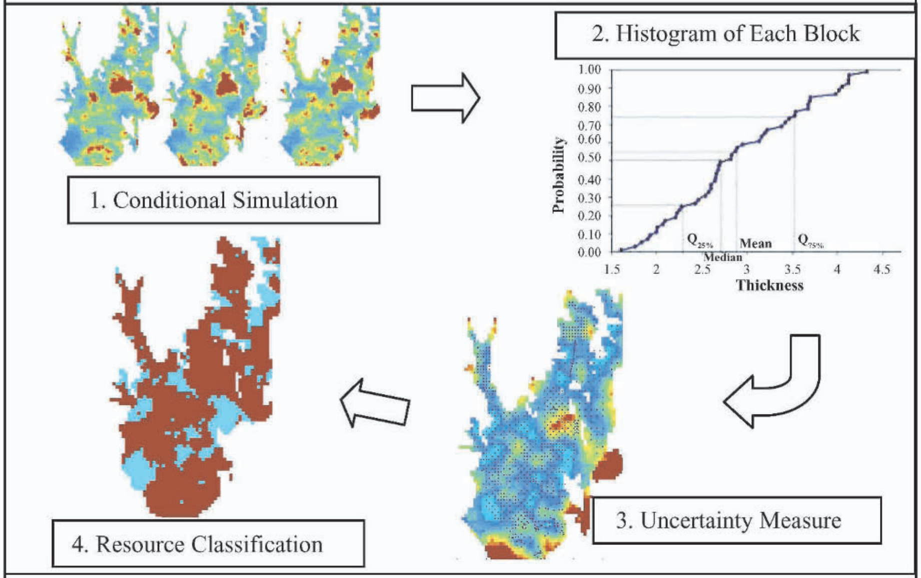


Figure 4 : Illustration of methodology

4. Conditional Simulation

The conditional simulation of resource thickness for the Grey Beds and Brown Beds at Elizabeth Bay was carried out using a Turning Bands simulation algorithm (Lantuéjoul, 2002). One hundred realisations were generated with a regular discretisation of 10 x 10 x 1. A total of 920 samples were used for simulating the Grey Beds and 727 for the Brown Beds. The simulation domain was defined by a polygon delineating the edge of the resource and includes 2882 blocks for the Grey Beds and 2910 blocks for the Brown Beds. Statistics of the back transformed data compared favourably with the raw data and variograms obtained from point conditional simulation confirmed the accuracy of the simulations. The simulation mean (e-type) is plotted on the upper left of Figure 5 and is compared to five realizations. The e-type estimate shows a greater amount of smoothing than OK (Figure 3) and the individual realizations show high variability.

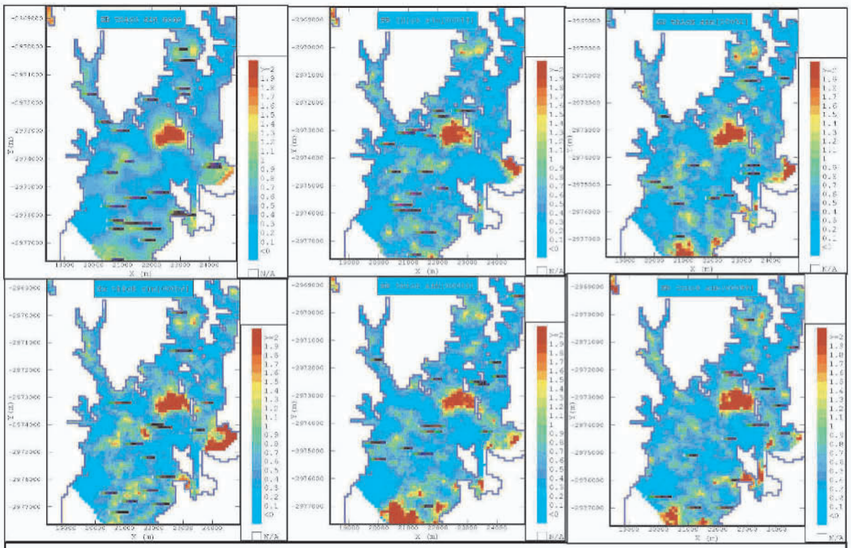


Figure 5: Simulation mean (e-type) for Grey Beds (top left) and five realisations

5. Measures of Uncertainty

The steps of conditional simulation and change of support are followed by the derivation of a set of conditional distribution functions (cdf) related to the simulated variable denoted by Z. Each of these functions provides a measure of local uncertainty in that it relates to a block attribute Z(u) at a specific location u within the deposit. From the distribution it is possible to read the probability that Z(u) is valued above any given cut-off z_k :

$$F(\mathrm{u};z) = P\{Z(\mathrm{u}) \le z_k\}, \quad \forall z$$

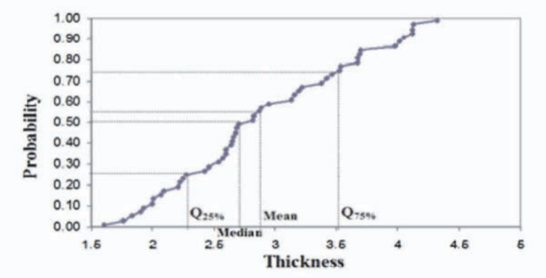


Figure 6: Example of a local conditional distribution function (cdf) for thickness at a given block

Figure 6 shows a cumulative conditional distribution function modelling the possible uncertainty about the thickness for a given block at location u. Each discrete point in the cdf corresponds to a simulated value z^l(u). The graph is a cumulative histogram containing all simulated values assigned to the block by each one of the realisations. A continuous function is interpolated between the discrete points to enable the assessment of probabilities for any cdf value.

A variate of summary statistics and uncertainty measures can be derived from the cdf and used to support decision making. In this study a number of basic measures are investigated and these include a conditional variance, a conditional coefficient of variation (relative standard deviation), the interquartile range, and a probability interval.

5.1 CONDITIONAL VARIANCE

The conditional variance measures the spread of the cdf around its mean value z_E^* :

$$CVar(u) = \sum_{k=1}^{k+1} \left[\bar{z}_k - z_E(u)\right]^2 \cdot \left[F(u;z_k) - F(u;z_{k-1})\right]$$

where:

- z_k, k=1,…K, are K threshold values discretising the range of variation of z values
- $\bar{z}_k$ is the mean of the class z_{k-1}, (z_{k-1},z_k) which in case of a within class linear interpolation model corresponds to $z_k = \left(z_{k\text{-}1} + z_k\right)/2$
- $z_E^*\left(u\right)$ is the expected value of the cdf approximated by the discrete sum:

$$z_E^*(u) = \sum_{k=1}^{K+1} \bar{z}_k \cdot \left[F(u;z_k) - F(u;z_{k-1})\right]$$

5.2 CONDITIONAL COEFFICIENT OF VARIATION

The conditional coefficient of variation (CCV) or relative conditional standard deviation corresponds to the conditional standard deviation divided by the mean. The CCV expresses variability as a percentage of the mean, and is calculated as follows:

$$CCV\ (u) = \frac{\sqrt{\sum_{k=1}^{k+1} \left[\bar{z}_k - z_E(u)\right]^2 \cdot \left[F(u;z_k) - F(u;z_{k-1})\right]}}{z_E^*(u)}$$

5.3 CONDITIONAL INTERQUARTILE RANGE

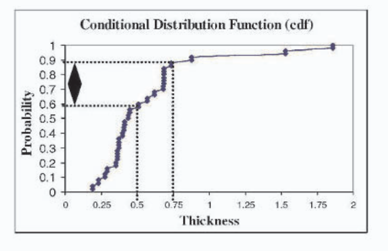
The conditional interquartile range (IQR) is defined as the difference between the upper and the lower quartiles of the distribution:

$$IQR(u) = q_{0.75}\ (u) - q_{0.25}(u) = F^{-1}(u;0.75) - F^{-1}(u;0.25)$$

5.4 PROBABILITY INTERVAL

The probability that the unknown is valued within an interval (a, b), termed probability interval, is calculated as the difference between cdf values for thresholds b and a:

$$\Pr ob\{Z(u) \in (a,b)\} = [F(u;b) - F(u;a)]$$



6. Results

The conditional variance (Figure 7) shows a poor correlation with sample density because there are high and low values related to both high and low sample density. However, the thicker part of the Grey Beds corresponds to a higher conditional variance value. Visual examination suggests there is a proportional effect between the conditional variance and thickness. The implication is that the conditional variance may not be an appropriate measure of local uncertainty if the objective includes the comparison between zones with different magnitudes of thickness.

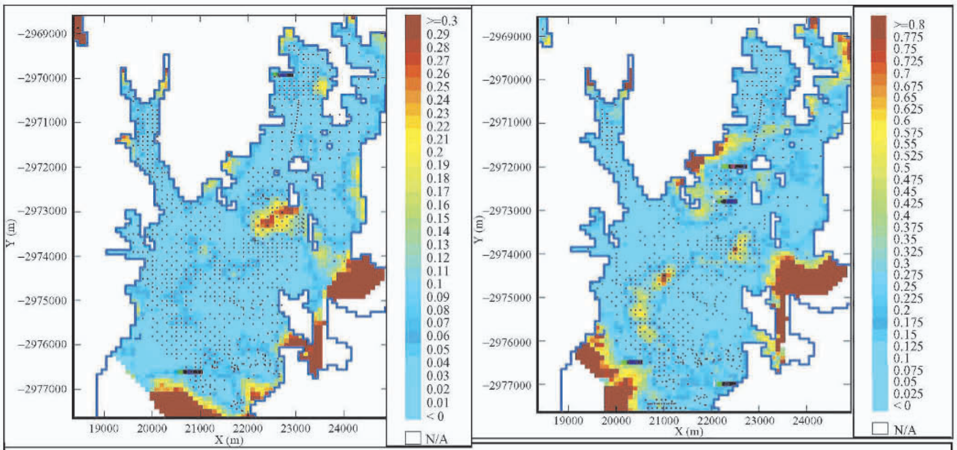


Figure 7 : Conditional variance for thickness; Grey Beds (left) and Brown Beds (right)

Figure 8 (left and middle) illustrates CCV values calculated for each block for the Grey and Brown Beds. There is a good correlation between CCV and conditioned sample points (points in Figure 8) and the identification of similar zones is better than with the conditional variance. Comparison with the estimation standard deviation (Figure 3) shows a difference. A further advantage of the CCV is that it is expressed as a percentage of the mean. The CCV values for the two horizons differ and that will make the selection of common threshold values for resource categories difficult.

The interquartile range (IQR) is a relative measure that does not use the mean as the centre of the distribution. This measure ignores the internal distribution of probability densities, leading to over representing uncertainty. Visual examination of IQR values for both horizons suggests higher uncertainty about the higher thickness values. The effectiveness of the IQR as an aid in classification may be improved by dividing by a median or mean to make the measure dimensionless (Lantuéjoul, 2003). The IQR divided by the mean for each block (Figure 8, right) shows an irregular change from low to high values making identification and selection of thresholds difficult. The major drawback of using IQR is that it ignores the distribution of probability densities, leading to an over representation of uncertainty.

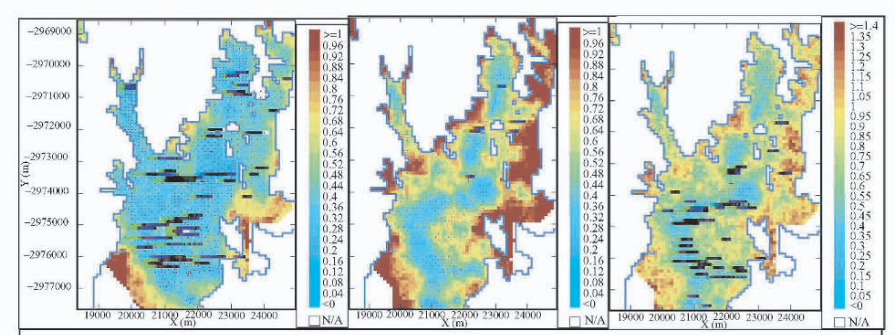


Figure 8 : Conditional coefficient of variation for Grey Beds (left), Brown Beds (middle) and IQR (right)

Probability intervals for a range of thicknesses at Elizabeth Bay could be selected between, for example, 0.5m and 0.75m. Like some of the previous measures, there is a good correlation between regions of dense sampling and high probability values. Where the resource thickness is less than 0.5m a less costly mining method can be used and this measure is ideal for establishing probability of finding a thickness of 0.5m or less in 100m x 100m blocks. However the measure is less useful for determining thresholds related to uncertainty.

7. Resource Classification

The resource classification criterion is based on the uncertainty measures derived from the cdf of each block in the resource model. Each criterion requires the selection of a threshold value that reflects the error tolerance that is acceptable for the block estimate.

Consider the CCV as a classification criterion. Given a threshold, λ, the 100m block will be classified

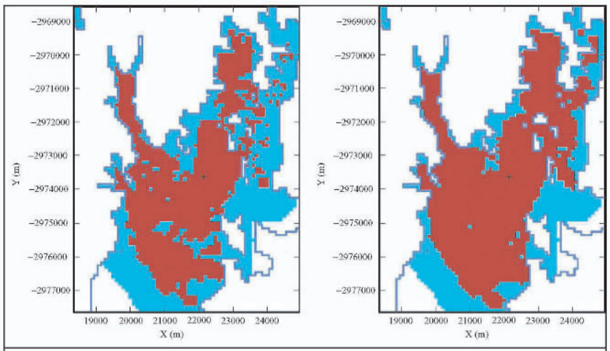


Figure 9 : Threshold values of 40% (left) and 50% (right) for the Grey Beds

at an Indicated or Inferred level of confidence depending on whether the CCV is less than or greater than the threshold; If CCV< λ then Indicated and if CCV $\geq \lambda$ then the block is Inferred. Figure 9 shows threshold values of λ =40% and 50% for the Grey Bed horizon where the central darker coloured blocks will be at an Indicated resource category.

8. Assessment of Sample Spacing Efficiency

The method, outlined in Figure 10 uses conditional simulation to quantify the expected error (% average difference between estimated and "actual"). The steps include initial analysis of the data, generation of a suitable conditional simulation, sampling each realisation using a given spacing, estimation of the attribute, calculating the expected difference error (%), repeat the process to generate expected errors and finally calculate the mean coefficient of variation of expected differences per block. The method was applied using data from a part of the Grey Beds at Elizabeth Bay. The deposit was simulated on a 10m x10m grid and node values, selected at a specific spacing were used to estimate thickness in 100m x 100m blocks. Fifty realisations using the sequential Gaussian simulation method (e.g., Dimitrakopoulos and Luo, 2004) were used to assess sample spacing at 50m, 100m, 150m, 200m and 250m intervals.

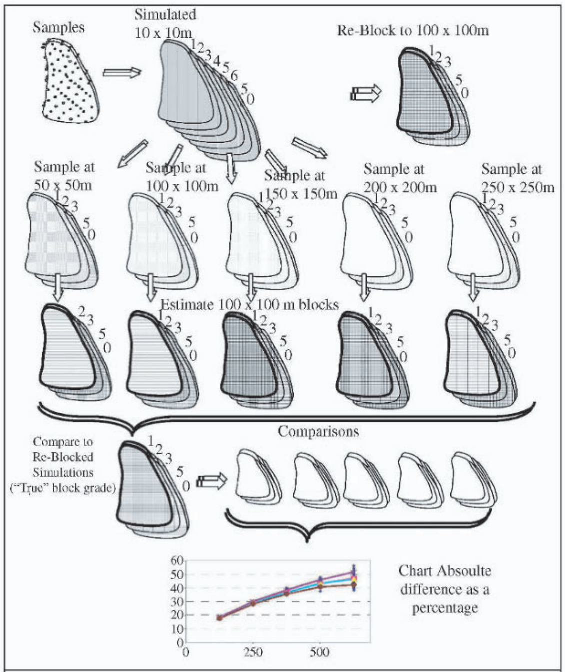


Figure 10: Illustration of method for finding optimum sample spacing

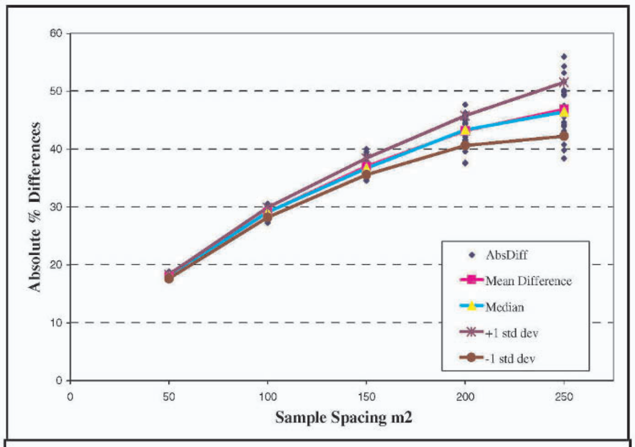


Figure 11: Expected differences for each sample spacing

The results of sampling the Grey Bed simulations are shown in Figure 11. The spread of differences found with 50 realisations for

50m x 50m sampling is very small increasing marginally to 100m x 100m sampling. The percentage absolute difference is sensitive to low "actual" thickness values. The expected difference for 50m x 50m sampling is about 18%, increasing to 55% for 250m x 250m sampling.

9. Discussion of Results and Conclusions

The methods using conditional coefficient of variation calculations provide promising results for quantifying uncertainty related to resource thickness estimates at Elizabeth Bay. The application of specific threshold values requires further work but these methods provide a useful tool for resource classification. Although not strictly quantitative when the results are combined with similar values calculated for grade, stone size and revenue they will enable a good overall classification of the resource to be made.

Of the measures evaluated the CCV provides the most reliable measure to use for resource classification and although the method remains semi-quantitative it is possible to use a CCV value to assign a resource category for the Elizabeth Bay Beds. Table 1 shows resource categories assigned by using CCV values where the Inferred category is assumed to be larger than Indicated or Measured and has been subdivided into "upper" and "lower" units. The classes have been determined predominantly by visual means but taking cognizance of the method referred to above and knowledge of the deposit. Figure 12 illustrates the application of CCV categories (similar to Table 1) on the Grey Beds (left) and the Brown Beds (right).

Table 1: Proposed Resource Classification categories derived from CCV values of thickness at Elizabeth Bay

CCV		Classification
0	0.2	Measured
0.2	0.4	Indicated
0.4	0.6	High level Inferred
0.6	0.8	Low level Inferred
0.8	2	Not in resource (Deposit)

The decision to allocate blocks to either an Inferred or Indicated resource category at Elizabeth Bay remains the responsibility of the "competent person" but using the CCV as a measure provides a good, quantitative method to use as an aid, which with similar measures on other variables, will make quantitative classification possible.

The method of assessing sample spacing efficiency can be used to determine the distribution of % errors for each of the 100m x 100m estimation block, and the mean errors mapped for each of the nominated sample spacings. The absolute average of expected % error in estimated thickness ranged from 18% at a spacing of 50 x 50m up to 55% for samples spaced at 250 x 250m.

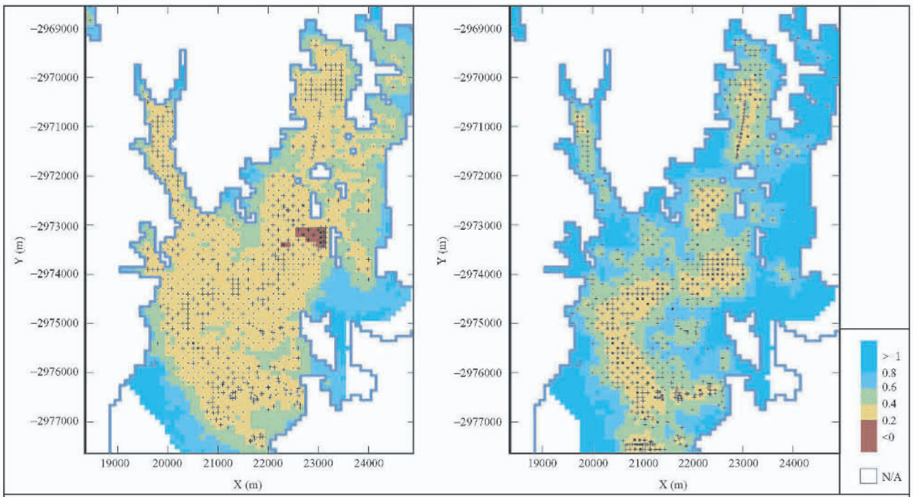


Figure 12: Final Classification of resources according to figures shown in Table 1 for the Grey Beds (left) and the Brown Beds (right)

Acknowledgements

Namdeb Diamond Corporation is acknowledged for allowing the presentation at Geostats 2004 and publication of this information. Dr. Marcello Godoy is acknowledged for his contribution to this study and Dr. Christian Lantuéjoul is thanked for his informative comments and discussion.

References

Dimitrakopoulos, R. and Luo, X., 2004. *Generalized sequential Gaussian simulation on group size v and screen effect approximations for large field simulations*. Mathematical Geology, v. 36, no. 5, pp. 567-591.

Duggan, S. P. 2002. *Mineral Resource Estimate for two portions of Elizabeth Bay.* Internal De Beers report written for Namdeb, November.

Godoy, M., 2003. *The effective management of geological risk in long-term production scheduling of open pit mines*. PhD Thesis, The University of Queensland, Brisbane, 252p.

Lantuéjoul, C. *Geostatistical Simulation.* Berlin: Springer, 2002, 192p.

Lantuéjoul, C. *Comparison of various criteria for resource classification.* Report written for De Beers, October, 2003.

GEOSTATISTICAL SIMULATION TECHNIQUES APPLIED TO KIMBERLITE OREBODIES AND RISK ASSESSMENT OF SAMPLING STRATEGIES

JACQUES DERAISME[1] and DAVID FARROW[2]
[1]*Geovariances, 49 Ave Franklin Roosevelt, Avon, 77212 France Deraisme@geovariances.com*
[2]*MRM - TSS, De Beers Consolidated Mines Ltd, P Bag X01, Southdale 2135 South Africa david.farrow@debeersgroup.com*

Abstract. Typically a kimberlite diatreme has several different geological zones. The upper portion is generally filled with the sedimentary crater facies, the central zone is more typically an in situ massive series of volcanic breccias and the lower regions comprise a complex root zone. Depending on the local degree of erosion, not all zones remain at any particular kimberlite occurrence.

A method of simulating the simpler internal geologies seen in the central region had previously been developed using a geometrical technique. In the upper reaches of the diatreme zone, the geologies have more complicated geometries and the approach adopted for the central regions needs to incorporate a more sophisticated method of simulating the internal geologies.

The similarity between the sedimentary facies that comprise the crater zone infill and the sequences that the oil industry targets as oil reservoirs suggest a similar technique could be applied to the simulation of internal geology of crater zone of kimberlite pipes.

Previous work has shown that a truncated gaussian approach can be useful, but the restrictions on facies relationships have limited its implementation. Plurigaussian simulation allows more complex interrelationships to exist between the simulated zones.

In conjunction with other geometric simulations, plurigaussian simulation can be used to guide sampling programs to optimise sampling layouts and sample size and ensure that the goals of the sampling programs are attainable. This paper focuses on the application of the combination of these simulation techniques and will be illustrated by a case study.

1 Introduction

The Orapa Kimberlite Mine forms the basis for this study. The mine, located 240 km west of Francistown in the western portion of the Botswana Central district, produces approximately 6 Million Carats per year with a value of almost US$ 500 million.

O. Leauangthong and C. V. Deutsch (eds.), Geostatics Banff 2004, 429-438.

Typically kimberlite pipes have a number of differing zones or facies. Three facies, namely "crater", "diatreme" and "hypabyssal" have been recognised in the Orapa orebody. The Orapa orebody comprises two volcanic pipes, which coalesce to form a single crater. The focus of this study is the crater facies rocks of the upper portion of the southern pipe.

The crater facies comprise a predominantly sedimentary type sequence of volcanic materials that have been re-deposited into the volcanic crater. Application of the Plurigaussian Simulation technique to the crater facies has been investigated for assisting with the creation of geological block models and determining an optimal sampling configuration.

Numerous boreholes and pit exposures have been combined into a digital geological model using GEMCOM™ software for analysis and visualization, that is typically a time consuming process. During mining, and as additional drilling is undertaken, more data becomes available. Incorporating additional data into the digital model requires that the models be regenerated, to ensure that the spatial distributions are maintained, once again a time consuming process. As a consequence, the digital models are not updated with any regularity and the mining models and the geological models are frequently out of phase. This results in sub-optimal Resource Management. An algorithmic method of more rapidly generating an overall geological model which can be updated on a regular basis is a significant advantage.

The initial search for a suitable algorithm explored the truncated gaussian approach. Despite showing promise, it did not prove very effective. The plurigaussian simulation methodology implemented in the latest release of the geostatistical software, ISATIS™, was another option which offers enhanced capabilities for geological modelling and has been successfully applied to the geological simulation of oil reservoirs. The application of this method is the subject of this paper.

2 The Geology of Orapa

A review of the Orapa geology is given in Field et al. (1997) and readers are referred to this paper for a more detailed introduction. For the purposes of this study only the major rock types of the crater facies are summarised.

The Orapa pipes intrude into the Archean basement granite-gneiss and tonalities and the sedimentary rocks and lavas of the Karoo Supergroup. They were covered by extensive thicknesses of Cenozoic and Mesozoic deposits. The deposit comprises two pipes, named the southern and the northern lobes. Rocks belonging to the crater, diatreme and hypabyssal facies, as described by Hawthorne (1975), have been recognised.

The Crater Facies deposits are well preserved and divisible into epiclastic, volcaniclastic and pyroclastic varieties. The epiclastic deposits are those in which sedimentary processes can be identified and comprise a wide variety of types including talus deposits, debris flow material, boulder beds, grits and lacustrine shales. Those deposits with no obvious mechanism of deposition are termed volcaniclastic. They are highly variable in character, with well sorted bedded horizons but dominated by coarse massive, matrix supported types. No convincing directional sedimentary structures have been found within the deposit. Basal Hetrolithic Breccias which apparently mark the base of the crater zone deposits occur intermixed with the volcaniclastic deposits. Pyroclastic deposits show evidence of direct deposition by volcanic mechanisms. The Pyroclastic deposits are present only in the northern pipe and comprise materials that exhibit evidence of pyroclastic fall, flow or surge.

3 The Plurigaussian Simulation Methodology

Plurigaussian simulation (PGS) aims to simulate categorical variables, such as geological facies, by the intermediate simulation of two continuous Gaussian variables. Facies are obtained by applying thresholds to the Gaussian simulated values. A detailed review of the PGS method is given in Armstrong et al. (2003).

The basic idea is to start out by simulating at grid locations one (Truncated Gaussian Simulation or TGS) or two (PGS) gaussian variables with a variogram characterizing the spatial continuity of the lithotypes indicators. Then a “rock type rule” is used to convert these values into lithotypes. The conversion is using the bijection between the gaussian values and the cumulated distribution function (cdf.). Therefore the application of that method requires to inform each grid node by the an estimate of the cdf. This step is carried out by calculating the so-called vertical proportion curves. By interpolating these proportions on the 3D grid, we get a 3D matrix of proportions.

PGS is an extension of TGS, the latter implying a rather strict stratigraphic sequence: because the simulated Gaussian values are continuous, the application of a threshold practically means, in the simple case of 3 facies, that for going from facies 1 to facies 3, it is likely to have a transition through facies 2. By using 2 gaussian functions, all transitions facies 1 to 2 or 1 to 3 or 2 to 3 are authorized.

The concept of non stationary proportion curves is central in PGS/TGS, where the so-called rock type rule plays an essential role in producing realistic models that represent the transitions between the different facies. The key point is that the Gaussian variables and the indicators are linked by means of thresholds but, even if the indicators are not stationary, they can be obtained by truncation of stationary Gaussian variables, which can be easily simulated. Initial applications were made in the petroleum industry where this approach seems natural due to the sedimentary origin of the reservoirs. The analogies with orebodies where mineralization occurs in layers forming a consistent stratigraphy justifies the application of the same conceptual model in this case study.

The process of PGS has three steps:

- determination of the vertical proportion curves from statistics on the drill-hole data. A vertical proportion curve represents the profile along the vertical of the proportions of each facies level by level. These statistics are highly dependent on the choice of a particular surface, the reference surface, which can be interpreted as a guide to the system of deposition of the different lithological facies. The drillhole data will then be transformed into a “flattened” space where the reference surface represents the horizontal surface at zero elevation. The simulations of the Gaussian variables will be processed in the flat space before being transferred to the real stratigraphic space.

- choice of a model describing the relationships between the different facies. This includes the definition of the lithotype rule, the correlation between the two Gaussian variables and their variogram models.
- generation of gaussian values at data locations. This is the most difficult and original part of the method, because at the data locations only facies are known but this does not tell the corresponding gaussian values. A special statistical method called a Gibbs sampler is used to generate these values.
- simulation of the two Gaussian variables followed by truncation to obtain the facies indicators. Finally the simulated facies are transferred to the structural grid.

4 Data Sources

o

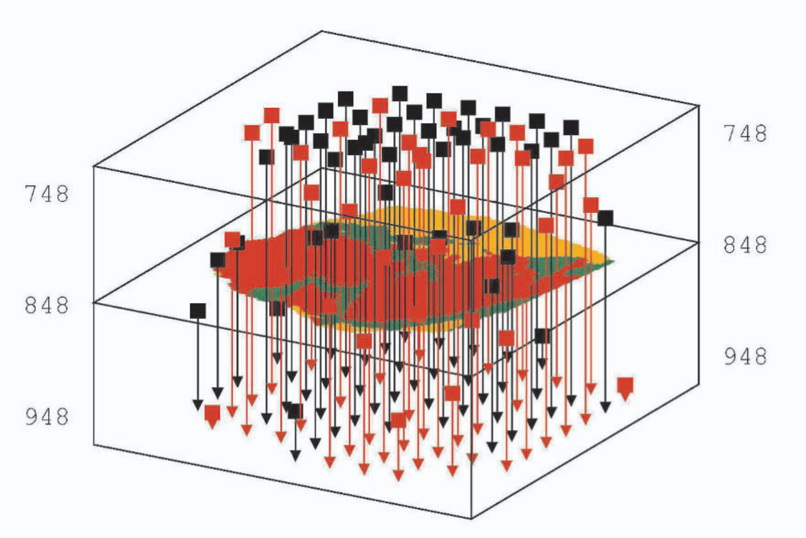


Figure 1: Vertical boreholes spaced on a 100m square grid, sub-sampled 200m grid boreholes in red.

The most recent geological model was transferred to a 5m * 5m * 5m block model. This was taken as the starting point of the work. Simulated boreholes were generated from the geological block model on 100 m, 150m and 200 m square grids (Figure 1). The facies observed on the simulated boreholes from the geological model are considered to be “reality” and are used to condition the facies simulations. The aim is to determine how much drilling is required to produce an accurate model of the pipe geology and associated volumes.

5 The Simulation

5.1 CHOICE OF A REFERENCE SURFACE

This is a crucial decision that has consequences on all stages of the process, data analysis and simulated images. In a sedimentary context the reference surface is meant to represent the direction perpendicular to the deposition of the different facies. When comparing facies parallel to that surface, more similarity is expected and consequently more correlation between boreholes than along parallel plane surfaces will be observed. The consequence on the simulated images will be to force the facies to be stacked in parallel to the reference surface. In the present case, a bowl shaped surface showing the angle of dip of the bedded horizons in accordance with the proximity to the pipe boundaries was used.

5.2 VERTICAL PROPORTION CURVES

The boreholes were discretized by cores of 5m in length and repositioned relative to the reference surface. For each case corresponding to the different horizontal spacing the vertical proportion curves have been calculated and averaged within polygons designed in order to take account of the lateral facies change (Figure 2).

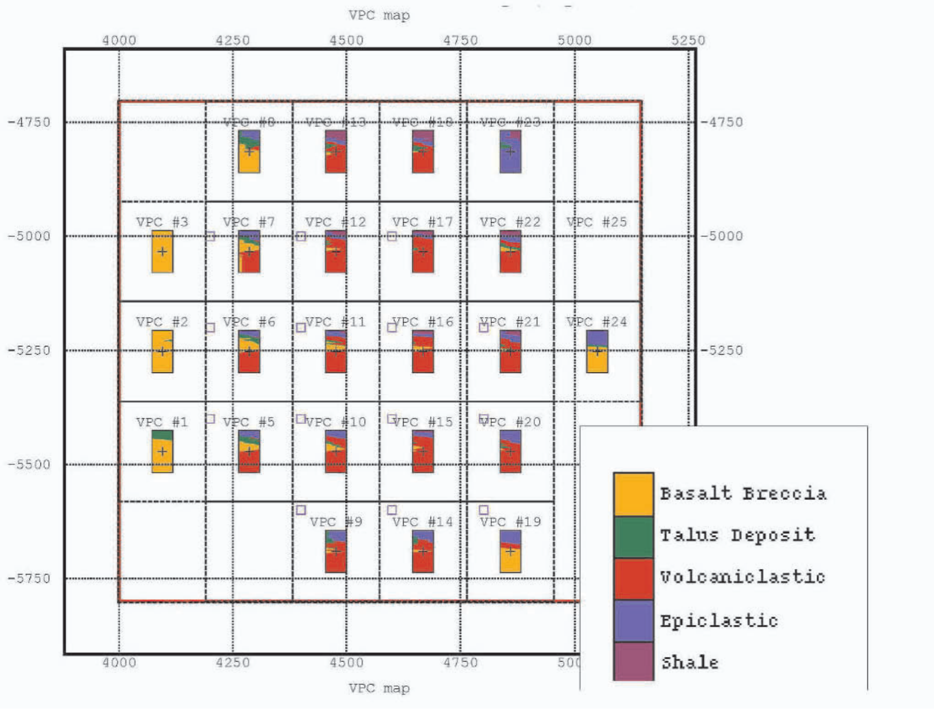


Figure 2: Vertical proportion curves calculated from boreholes within 2D polygons.

5.3 3D PROPORTIONS

The vertical proportion curves have then been interpolated on each grid cell by a kriging procedure with a rather long range (2 km) variogram, expressing the very gradual change of the lithotype proportions. This gives a 3D matrix of proportions that will be used to calculate local thresholds on the Gaussian random functions (Figure 3).

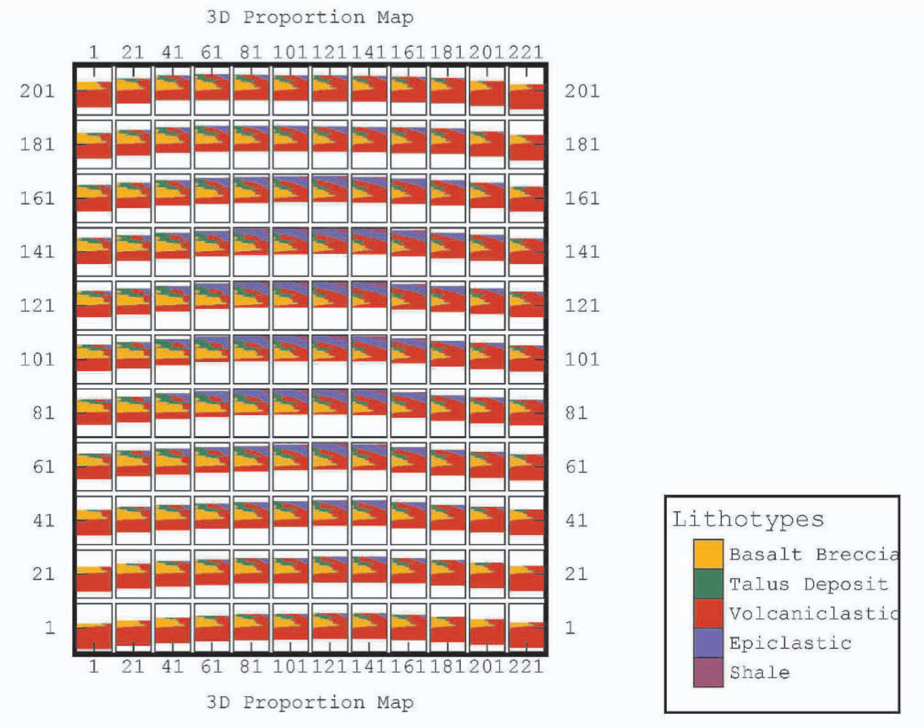


Figure 3: 2D representation of the 3D proportions interpolated on the grid in the flat space.

5.4 LITHOTYPE RULE

The knowledge of the lithotype proportions is not sufficient to derive the values of the thresholds to be applied on the simulated Gaussian values. Additional information on a partition of the 2D gaussian space is required. Depending on the number of facies, there is a finite number of rectangular partitions that may represent the possible relationships between the Gaussian random functions and the lithotypes. From these we chose the most sensible from a geological point of view, considering the probable transitions between the facies. For instance, since the shale occurs on the top of the diatreme adjacent to the epiclastic deposit, while the basalt breccia mark the base of the crater, it is appropriate to differentiate the corresponding lithotypes on the first Gaussian function. The representation (Figure 5) is schematic: the thresholds will be calculated at the simulation stage. In this example the lithotype rule implies that Basalt, Epiclastic and Shale are dependent only on thresholds applied on the first Gaussian function, while the Talus and Volcaniclastic also depend on the second Gaussian function.

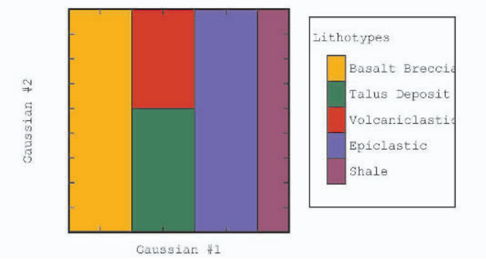


Figure 4: Rectangle lithotype rule.

Once the lithotype rule is defined, the variograms of the two Gaussian functions can be modelled. The prevailing role played by the proportion curves does not mean that the choice of the variogram has no consequence. This is illustrated in the Figure 5, where two ranges of the variogram of the first Gaussian function were tried as an example. In the lower picture the 3 lithotypes (orange,blue and purple), that are only discriminated by the first Gaussian function look much less continuous than on the upper picture. The second Gaussian function was simulated in correlation with the first (coefficient of correlation of 0.7) in order to make the Talus facies (green) preferentially conformable to the Breccia facies (orange).

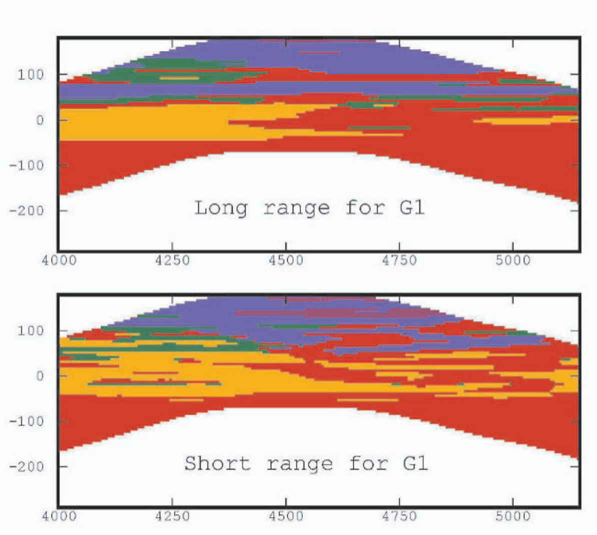


Figure 5: Cross section (in the flat space system) of two simulations changing the horizontal range of the variogram associated to the first Gaussian function.

5.5 CONDITIONAL SIMULATIONS

The simulations, achieved by means of the turning bands method, are performed in flat space, then transferred to the real “stratigraphic” space. Figure 6 compares the original geological model to 3 different realizations obtained from plurigaussian simulations using either no boreholes (just the average proportions) and called “non conditional” or boreholes (BH) spaced every 200m or every 100m. It is observed that with an increasing availability of data, the simulations converge towards the supposed reality.

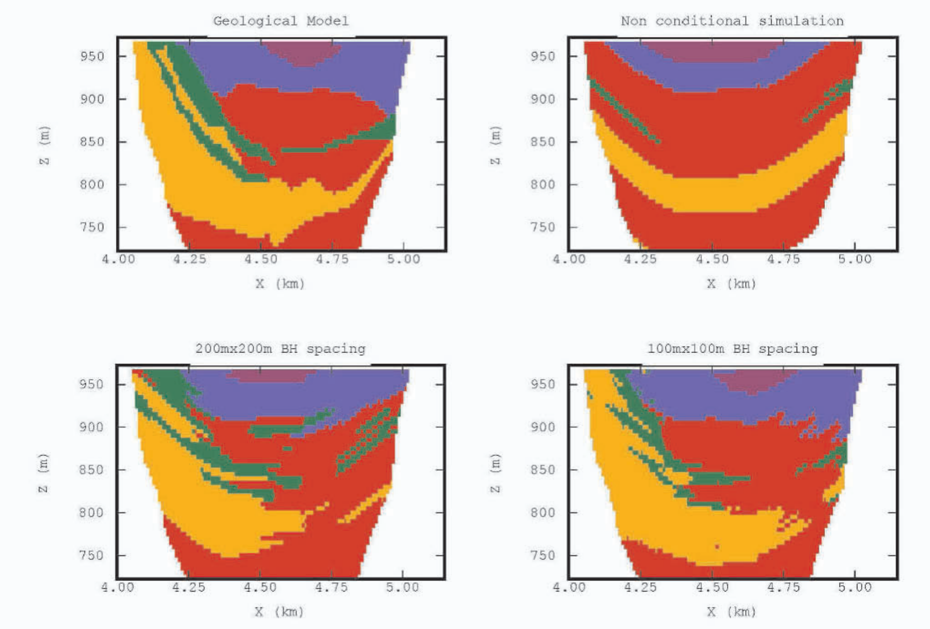


Figure 6: Cross section of three plurigaussian simulations with increasing information rates compared to the original block model.

6 Results

In the scope of evaluating the level of uncertainty in the volume of each lithotype, statistics have been calculated on the difference between the original block model and the simulations based on different levels of information. By comparing different borehole spacings (100m, 150m and 200m) it appears that 150m provides a satisfactory global estimation of the different facies (Figure 7 where the density of boreholes has been transformed into metres drilled).

The detailed analysis of the volumes of the different lithotypes has been made by levels 25m high (Figures 8 and 9). Compared to the geological model, it appears that the sampling using boreholes on 100m spacing guarantees maximum reduction in uncertainty. Use of boreholes on 200m spacing boreholes leads to an average uncertainty of about 10%, rising to 20% for some levels.

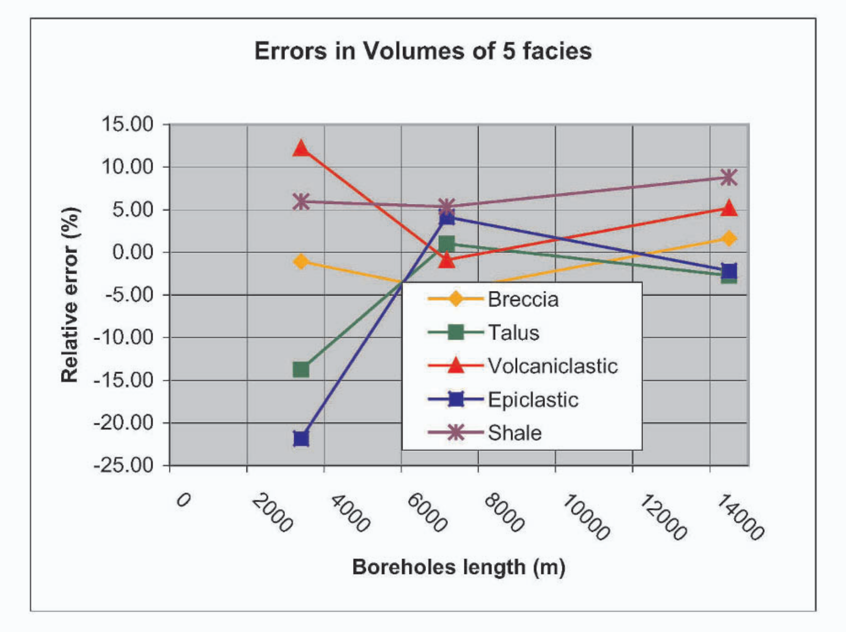


***Figure** 7*: Relative errors on the global volumes from simulations preformed with increasing sampling by boreholes.

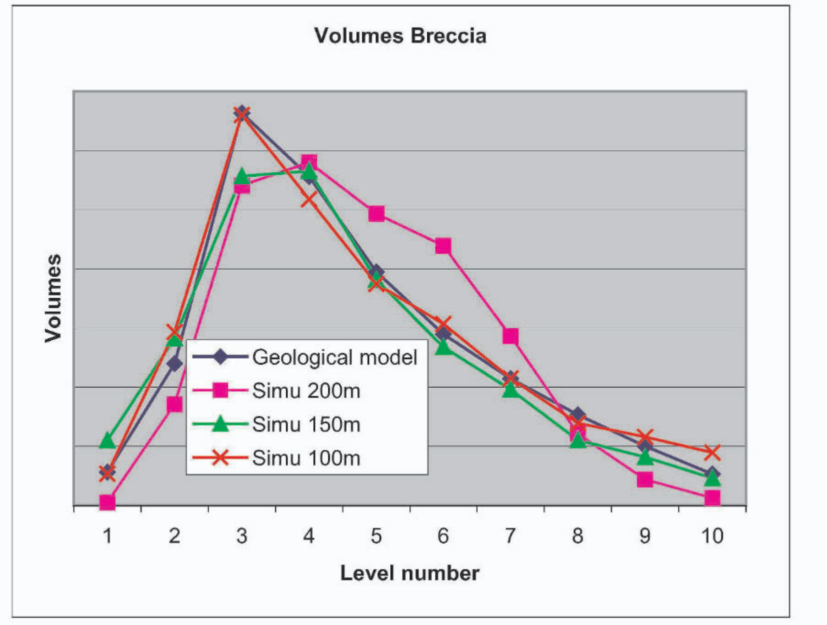


***Figure** 8*: Volumes of Breccia on 25m high levels, of the geological model and the simulations with different sampling

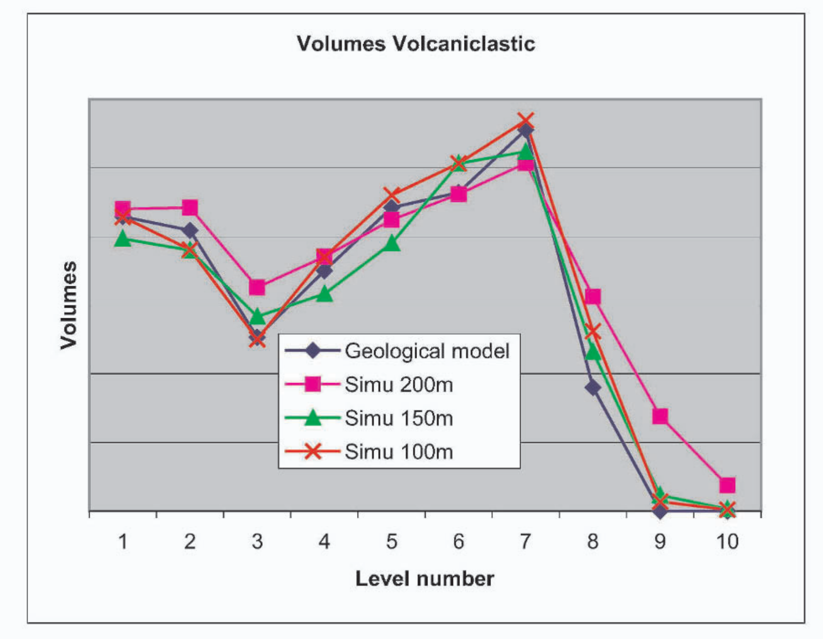


Figure 9: Volumes of Volcaniclastic, on 25m high levels, of the geological model and the simulations with different sampling.

7 Conclusions

The plurigaussian simulation process has proved to be very efficient in providing images reproducing the main features of the geology encountered in kimberlite crater deposits. It appears this may be a useful addition to the process of geological modelling. This will be explored in an operational context. Besides, the quantification of the confidence as a function of the number of holes will aid in economic decision making.

References

Hawthorne, J.B.. Model of a kimberlite pipe. Physics and chemistry of the earth, 1975, Vol 9, p.1-15.

Armstrong A., Galli A., Le Loc'h G., Geffroy G. and Eschard R., *Plurigaussian Simulations in Geosciences*, Springer, 2003.

Bleines. B., Deraisme J., Geffroy F . et al., *Isatis software Manual*, Géovariances and Ecole des Mines de Paris, 2004.

Chilès J. P. and Delfiner P., *Geostatistics Modeling Spatial Uncertainty*, Wiley & Sons, 1999, p. 531-535.

Deraisme J. and Farrow D., *Quantification of uncertainties in geological modelling of kimberlite pipes*, APCOM 2003, The South African Institute of Mining and Metallurgy, 2003.

Field, M., Gibson, J.G., Wilkes, T.A., Gababotse, J. and Khutjwe, P. (1997). The geology of the Orapa A/K1 kimberlite Botswana: Further insight into the emplacement of kimberlite pipes. Russian Geology and Geophysics, Vol.38, No.1. Proceedings of the Sixth International Kimberlite Conference, Vol. 1, p.24-41.

MODELLING 3D GRADE DISTRIBUTIONS ON THE TARKWA PALEOPLACER GOLD DEPOSIT, GHANA, AFRICA

THOMAS R. FISHER*, KADRI DAGDELEN**, and A. KEITH TURNER*
**Department of Geology and Geological Engineering*
***Department of Mining Engineering*
Colorado School of Mines
Golden, Colorado 80401-1887 USA

Abstract. In the Precambrian Tarkwaian Group of Ghana, gold is preferentially located as paleoplacers within quartz-pebble conglomerates. Gold distributions are intimately associated with sedimentologic and stratigraphic features of the host rocks. In this situation, traditional geostatistical methods have not provided accurate predictions of ore grades and reserves, due to difficulties in properly incorporating geologic information in the geostatistical estimation.

Application of Transition Probability/Markov geostatistical techniques allowed us to combine geologic concepts and domain knowledge with indicator and Gaussian-based estimation techniques. Vertical variability relationships within stratigraphic sequences, as measured by borehole data, were used to predict lateral distributions of lithologic facies. The result was a set of 3-D spatial relationships that reflect an integration of geologic concepts and readily observable geologic attributes.

This approach provides an alternative to more traditional geostatistical ore deposit modelling. It provides a statistically sound, lithofaces-based prediction of gold grades and uncertainty of the predictions, constrained by geology and the 3D geological framework.

1 Introduction

Mine profits are largely determined by accurate estimation of ore reserves and correct classification of material as either ore or waste during mining operations. Accurate prediction of ore grades and reserves requires incorporation of geological data, knowledge, expertise, and concepts. In the Paleoproterozoic Tarkwaian Banket Formation of Ghana, West Africa, gold is preferentially located in a succession of paleoplacer within quartz-pebble conglomerates. Thus, gold distributions are associated with sedimentologic and stratigraphic features of the host rocks.

At the Tarkwa mine, simple kriging based on 100m x 100m diamond drill (DD) boreholes underestimates gold values by as much as 20 percent below reported mill head grades (Gold Fields Ghana, Ltd, 2003). It is believed that this inaccurate

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 439-448.

prediction of ore grades may result from lack of characterization of depositional environments of the host rocks in estimation of ore reserves. Underestimation may not cause great difficulties during current opencast mining operations, but the success of projected future underground operations with less densely spaced borehole control will depend on much more precise predictions of both ore grades and reserves based on appropriate geologic models.

2 Regional Geologic Setting

The Tarkwa region (Figure 1) is contained within the Man-Leo Shield in southwestern Ghana. The geology is dominated by the Paleoproterozoic Birimian Supergroup, a series of meta-volcanic belts and intervening meta-sedimentary basins that formed as primitive island arcs accreted to the Archean craton (Sylvester and Attoh, 1992). The Birimian terrane consists of five NE-SW trending volcanic belts, named from east to west (and youngest to oldest): the Kibi-Winneba Belt, the Ashanti Belt, the Sefwi Belt, the Bui Belt, and the Bole-Navrongo Belt.

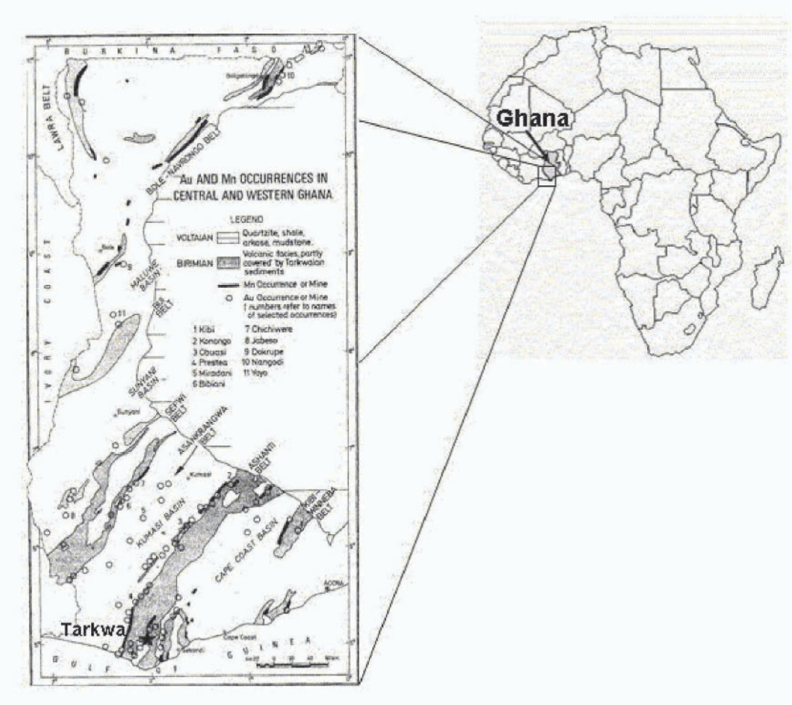


Figure 1. Index map and simplified geology of the Tarkwa Region (after Fisher and Turner, 2002).

The Tarkwa mine is located at the southwest end of the Ashanti Belt or "Tarkwa Syncline" (Figure 1). At least five episodes of deformation have affected the syncline. The Tarkwa depositional basin was formed during the first episode of deformation and is filled by a fining-upward sequence of Proterozoic clastic sedimentary rocks, known as the Tarkwaian Group, that is indicative of an extensional half-graben geometry as described by Frostick and Reid (1987). The Tarkwa Syncline contains the largest accumulation of Tarkwaian-type sediments and the highest paleoplacer gold concentrations of the Ashanti Belt. The underlying Birimian is considered a possible source of the gold in the Tarkwaian placers (Boadi, et al, 1991), although there is much disagreement amongst researchers regarding this theory.

The Tarkwaian Group consists of four formations, the Kawere, the Banket, the Tarkwa Phyllite, and Huni Quartzite. The Banket, main gold-bearing unit of the Tarkwaian, consists of up to 160m of relatively mature quartzites and conglomerates. Gold is found in paleoplacers within a succession of stacked, tabular units of alluvial fan, braided-stream, and valley-fill deposits derived from a southeastern source. At the Tarkwa Mine, the primary Banket gold concentrations are located in the A1, A3, and C zones. Gold is extracted from five areas of the mine; Pepe Anticline, Mantraim, Akontansi East, Akontansi Ridge, and Kottraverchy.

3 Modelling Procedure

An overview of the adopted modelling procedure (Fisher, 2004) is given in Figure 2. Geostatistical methods used incorporated site-specific geological information, knowledge, and experience to translate raw observations into a 3-D probabilistic model of gold distribution in the Pepe Anticline area of the mine. The top of Figure 2 shows the three major types of information incorporated within the process – exploration drilling data, field observations, and geologic maps and cross-sections. Gold Fields Ghana Ltd. provided much of these data in digital formats, but these historical data sources were supplemented by personal observations and discussions with mine personnel during an extensive site visit in early 2002. These data consisted of two major types – data associated with exploratory drilling that were used to create a borehole database (upper left of Figure 2), and stratigraphic and structural information (Box 4, Figure 2) that formed the basis of a 3-D geologic framework model of the Pepe area of the mine.

The modelling procedure has two major components. The first component, shown on the left side of Figure 2 (Boxes 1-3), involved statistical assessment of the observations contained in the borehole database to define statistically and geologically meaningful sedimentological units, herein called “statistical facies”, and subsequently the development of probabilistic distributions of gold-assay values for each “statistical facies”. The second component, shown on the right side of Figure 2 (Boxes 4-7), involved several steps to develop a 3-D probabilistic model of the spatial distribution of the “statistical facies”. The final step in the modelling process (shown as Box 8, Figure 2) produces a 3-D probabilistic model of the gold distribution by combining the 3-D probabilistic distribution of the “statistical facies” (produced from the second component and shown as Box 7, Figure 2) with the probabilistic distributions of gold-assay values for the “statistical facies” (produced from the first component and shown as Box 3, Figure 2). The production of a 3-D probabilistic model of the gold distribution permits several useful applications to mine planning and operation (Box 9, Figure 2). Details of this modelling procedure are provided in the following sections.

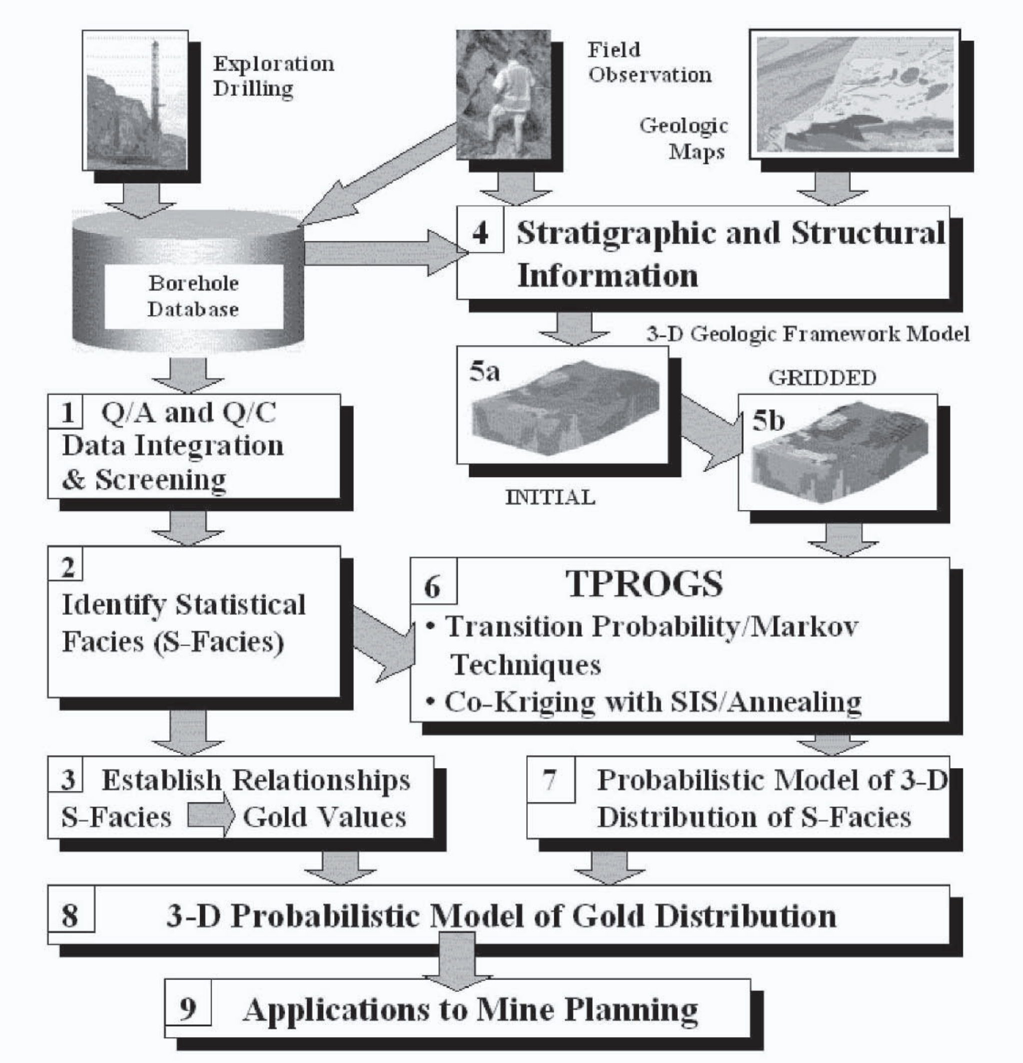

Exploration Drilling
Field Observation
Geologic Maps
Borehole Database
4 Stratigraphic and Structural Information
3-D Geologic Framework Model
5a
INITIAL
GRIDDED
5b
1 Q/A and Q/C Data Integration & Screening
2 Identify Statistical Facies (S-Facies)
6 TPROGS
• Transition Probability/Markov Techniques
• Co-Kriging with SIS/Annealing
3 Establish Relationships S-Facies Gold Values
7 Probabilistic Model of 3-D Distribution of S-Facies
8 3-D Probabilistic Model of Gold Distribution
9 Applications to Mine Planning

Figure 2. Flowchart of the Analysis Process (from Fisher, 2004).

3.1 AVAILABLE DATA

Data from 53 DD, continuously-cored exploration boreholes located on a nominal 100m by 100m grid within the Tarkwa mine were extracted from the main Gold Fields Ghana, Ltd. Mine databases. The critical observations were contained in several data sources. These had to be merged to form a single consistent borehole database for this study.

When completed, this database contained for each borehole: 3-D coordinates, stratigraphic and lithologic observations (including 19 categorical and continuous geologic variables) developed by geologists logging the cores, plus gold assay values obtained from core samples. An extensive QA/QC data-screening process was undertaken to ensure this master database was error-free and appropriately formatted for use in the study (Box 1, Figure 2).

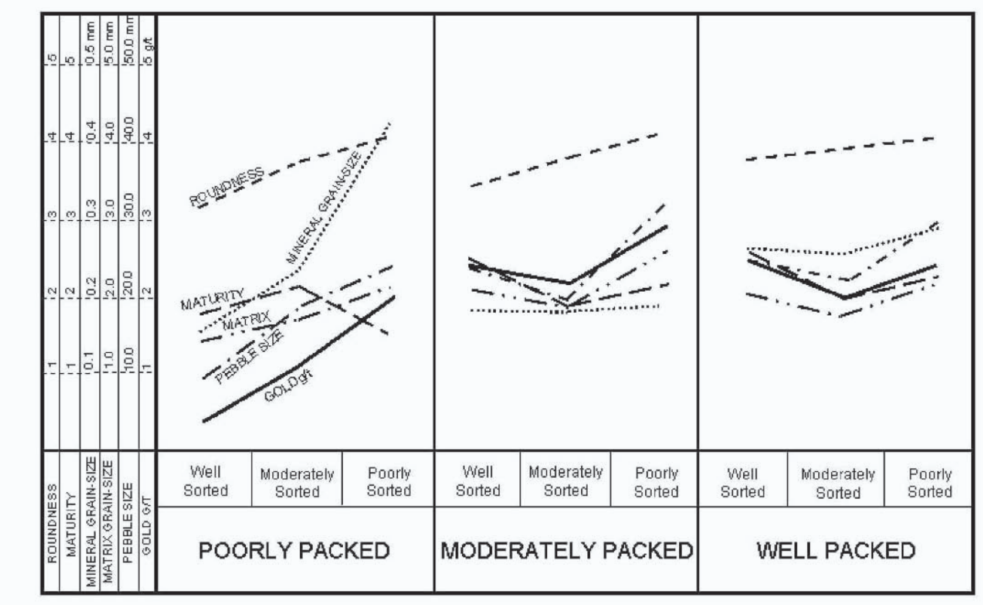


Figure 3. Relationships between gold values and selected sedimentological parameters for "sorting-packing" (pebble and larger-size material) pair groups. Values for each parameter are averaged from 528 core samples. Roundness refers to pebble or cobble roundness, mineral grain-size pertains to heavy mineral assemblage (primarily magnetite and hematite). Maturity and matrix refer to coarse sand-sized and smaller material.

3.2 IDENTIFICATION AND CREATION OF STATISTICAL FACIES

Studies of the Tarkwaian Banket by Sestini (1976) and Hirdes and Nunoo (1994) demonstrate that gold distributions reflect stratigraphic features, such as channel geometry and lithofacies, and are related to sedimentological parameters. Such relationships are geologically reasonable in placer deposits (Burton and Fralick, 2003). The master borehole database contained significantly greater numbers of observations than had been available in the earlier studies, so a series of correlations was computed between gold assay values and stratigraphic and sedimentological parameters. Although the data proved to be somewhat noisy, the conclusions reached by the earlier studies remained valid. Figure 3 shows gold values increasing as grain/pebble packing and sorting improves, and varying systematically with other selected sedimentological parameters. A k-means clustering method (Wishart, 2001) from ClustanGraphics® using a modified version of the Gower general similarity coefficient (Gower, 1971), permitting simultaneous handling of both categorical and continuous variables, was used to identify geologically meaningful groups from the parameters (Fisher, 2004). Six clusters were selected and designated as “statistical facies” (Box 2, Figure 2). The “statistical facies” can be related to real world conditions in a manner similar to the “electro-facies” concept of Doveton (1994a).

3.3 ESTABLISHMENT OF GOLD DISTRIBUTIONS FOR STATISTICAL FACIES

By coding the membership of each sample in the borehole database according to its “statistical facies,” and associating the samples with their proper assay values, it was possible to compute distinct gold distributions for each “statistical facies” (Box 3, Figure 2). These distributions form the basis of gold assay **cdf**’s for each “statistical facies”.

3.4 DEVELOPMENT OF 3D GEOLOGICAL FRAMEWORK

A 3D geologic framework was constructed to constrain the geostatistical simulations (Fisher, 2004). A series of correlated and tied cross-section panels was developed from the available 100m x 100m spaced DD borehole data. Tops and bases of principal gold-bearing Banket members A1, A3, and C were “picked” along with positions of the main thrust fault planes (Fisher, 2004). This interpretation provided opportunity to properly locate and correlate the main thrust faults and distinguish relationships within the three main fault blocks identified within the study area.

Stratigraphic horizons and fault planes were individually hand contoured. These maps were then digitized and cartographically registered in ESRI’s ArcMap® GIS software. The results were then exported to Golden Software’s SURFER® 8.0 contour and volume modeling software, where the digitized contours of each horizon were gridded to produce a series of individual 3D surfaces. Within SURFER® it was possible then to stack the constructed surfaces, including interpreted fault planes, into a 3D framework volume model (Box 5a, Figure 2 and Figure 4). This resultant model was then discretized into 5m by 5m (in the X-Y plane) by 0.5m (in the Z-direction or vertical direction) 3D cells (Box 5b, Figure 2) that were used by the T-PROGS transition probability software.

3.5 TRANSITION PROBABILITY MODELLING

Markov chain analysis forms the basis of the transition probability approach (Box 6, Figure 2). It has been successfully used in stratigraphy and sedimentology to discover statistically significant and fundamental patterns of lithological repetition (e.g., Doveton, 1994b). An initial 1D Markov model can be extended to a 3D Markov model to predict lateral distributions of sedimentary facies, where lateral variation has been under- sampled, by using Walther’s Law (Middleton, 1973) and changing the associated mean lengths in accordance with geological expectations and observed field data (Weissmann, et al, 1999). The T-PROGS software (Carle, 1999) was selected for this phase of the modeling because it permits incorporation of categorical variables and subjective observations into the model. The reader is referred to Carle (1996), Carle and Fogg (1996) and Weissmann, et al, (1999) for more in-depth explanation and details.

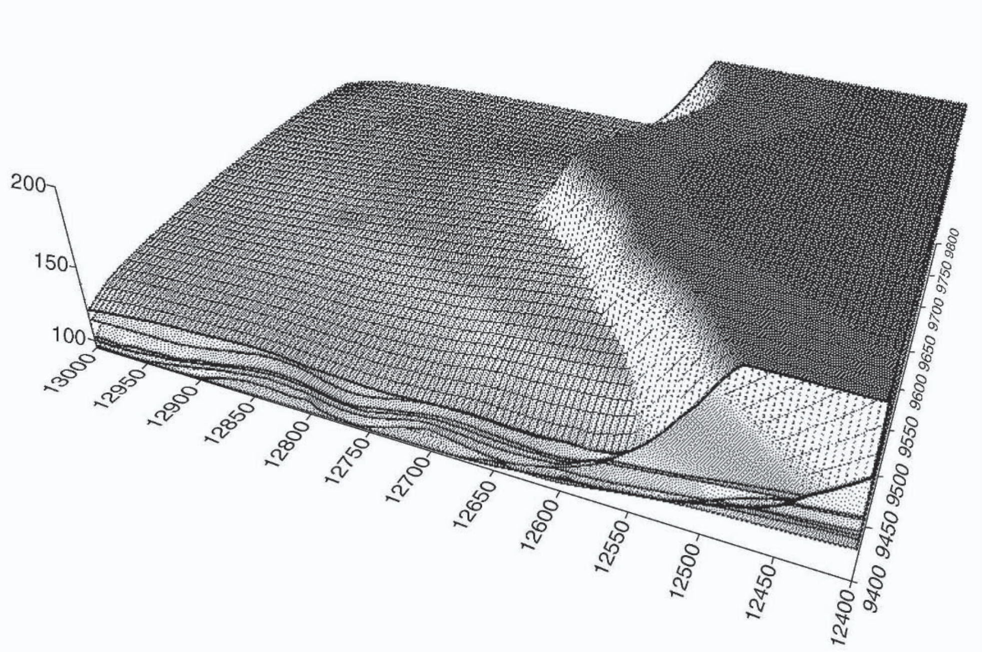


Figure 4. Display of initial 3D geologic framework model (Fisher, 2004).

3.6 SEQUENTIAL INDICATOR SIMULATION AND SIMULATED ANNEALING

Sequential Indicator Simulation, or "SIS", (Journel and Kyriakidis, 2004) was iteratively applied to the 3D Markov model transition probabilities defined in the previous section and substituted for the indicator cross-variogram in a co-kriging step (where the 3D Markov model conditioned the simulation) to produce 3D realizations of the modeled facies (Weissmann, et al, 1999). The SIS method was used because no assumptions are necessary about the shape of the gold distributions. The Markov chain model controlled factors such as global lithofacies proportions calculated from the conditioning borehole data and juxtapositional patterns. The co-kriging equations are solved using a basis function approach (Carle, 1996) - a more computationally efficient method. Subsequently, Sequential Quenching, the "zero-temperature" case of Simulated Annealing (van Laarhoven and Aarts, 1987), was applied to improve the geological reality of the SIS realization and reflect additional constraints (Fisher, 2004). The Sequential Annealing step improves the match of the final realization with the originally computed transition probabilities. The SIS realization was gradually perturbed so as to match defined characteristic lengths and facies continuities. No changes of facies were allowed at the boreholes, because these locations were "known", but at other locations a series of stochastic variations in facies assignment was possible.

Repetition of this process provided for the production of an uncertainty-based 3D distribution of "statistical facies" (Box 7, Figure 2).

3.7 DEVELOPING A 3D PROBABILISTIC MODEL OF GOLD DISTRIBUTION

Our ultimate goal was to produce a lithofacies-based gold grade uncertainty distribution (Box 8, Figure 2). This was accomplished, on a cell-by-cell basis, combining the uncertainty-based distribution of statistical facies (Box 7, Figure 2) with the previously computed **cdf** for the gold distribution of each "statistical facies" (Box 3, Figure 2). The computational process involves several steps. Because the probabilities of facies occurrences were altered during the quenching phase, the probabilities from the quenching step were used to compute a "global" **cdf** for all facies. Each model cell was examined in turn, and the most probable facies for each cell selected from this **cdf**. A facies having been selected, the appropriate **cdf** of gold distribution was selected and used to assign a gold grade to the cell. This process was repeated multiple times, producing several different gold grade estimates for each cell (Fisher, 2004).

All of the assigned grades were then used to compute a cumulative grade distribution, and uncertainties in grade, for the entire realization. This information was accumulated and presented as **pdf** plots. Values were assigned to appropriate selective mining units (SMUs). Mean grades were computed for the SMUs and the material in the SMU classified as either ore or waste according to pre-selected cutoff values. Thus, the process provides a statistically sound, lithofaces-based gold grade uncertainty prediction constrained by geology and the 3D geological framework.

4 Conclusions and Applications to Mine Planning and Operations

Construction of a statistically sound, lithofacies-based 3D gold-grade distribution model, which incorporates uncertainty in the predictions, supports several important mine planning and optimization applications. Efficiency of mine planning can be improved by using 3D geological models during exploration, using new borehole data "on-the-fly" as it becomes available.

By substituting differing geological concepts prior to the Sequential Indicator/Simulated Quenching steps, we can apply the process and methodology defined herein to orebodies other than sedimentologically controlled paleoplacers. For instance, Carlin-type deposits, layered mafic/ultramafic intrusions, massive sulfide, or other ore deposits may also be evaluated by applying appropriate 3D geological model(s).

Lastly, traditional ore control techniques classify material into ore or waste categories based on estimated average grade assuming *no uncertainty* exists on the estimated grades (Coşkun, 1997). Adding uncertainty to the estimates of grade, based on knowledge of geologic conditions, allows methodologies such as the loss function concept (Coşkun, 1997; Dagdelen and Coşkun, 1998) to be applied with greater confidence and accuracy.

Acknowledgements

We would like to acknowledge and thank Gold Fields Exploration, Inc., Denver, Colorado; Gold Fields Ghana, Ltd., Accra, Ghana, Africa; Gold Fields Ltd., South Africa, and the personnel of the Tarkwa Mine, Tarkwa, Ghana, for providing access to the Tarkwa Mine data, and for their generous funding of this research. We also thank the U.S.G.S. Water Resources Branch, Tucson, Arizona for funding modification and research of the T-PROGS Transition Probability Geostatistical Software.

SURFER® is a registered trademark of Golden Software, Inc., Golden, Colorado. ArcMap® is a registered trademark of ESRI, Redlands, California. ClustanGraphics® is a registered trademark of Clustan Limited, Edinburgh, Scotland.

References

Boadi, I. O., Norman, D. I., and Appiah, H., Source Terrane for Tarkwa Paleoplacer Deposit, Ghana, *in*, Pagel, M., and Leroy, J. L., (eds.), *Source, Transport, and Deposition of Metals*, Proceedings of the 25th Ann. SGA, Nancy France, Balkema Publishers, 1991, p. 641-648.

Burton, J. P., and Fralick, P. W., Depositional Placer Accumulations in Coarse-Grained Alluvial Braided River Systems, *Economic Geology*, vol. 98, no. 5, 2003, p. 985-1001.

Carle, S. F., *A Transition Probability-Based Approach to Geostatistical Characterization of Hydrostratigraphic Architecture*, Unpublished Ph.D. Dissertation, University of California, Davis, 1996, 233p.

Carle, S. F., *T-PROGS: Transition Probability Software, Version 2.1*, University of California, Davis – Hydrologic Sciences Graduate Group, 1999, 78p.

Carle, S. F., and Fogg, G. F., Transition Probability-Based Indicator Geostatistics, *Mathematical Geology,* vol. 28, no. 4, 1996, p. 453-476.

Coşkun, B., *Risk Quantified Ore Control*, Unpublished Master's Thesis, Colorado School of Mines, 1997, 119p.

Dagdelen, K., and Coşkun, B., Risk Quantified Ore Control in Open Pit Gold Mining, *in,* Basu, A. J., (ed.), *Computer Applications in the Minerals Industries International Symposium*, Australian Institute of Mining and Metallurgy, Publication Series No. 5/98, 1998, p. 9-12.

Doveton, J. H., *Geological Log Analysis Using Computer Methods*, AAPG Computer Methods In Geology, No. 2, 1994a, p. 169.

Doveton, J. H., Theory and Applications of Vertical Variability Measures from Markov Chain Analysis, *in,* Yarus, J. M., and Chambers, R. L., (eds.), *Stochastic Modeling and Geostatistics – Principles, Methods, and Case Studies*, AAPG Computer Applications in Geology, No. 3, 1994b, p. 55-64.

Fisher, T. R., *Three-Dimensional Sedimentological and Geostatistical Modelling in Precambrian Paleoplacers of the Tarkwa District, Ghana*, Unpublished Ph.D. Dissertation, Colorado School of Mines, 2004, (in preparation).

Fisher, T. R., and Turner, A. K., Application of hybrid 3D modelling methods to prediction of ore grades in stratabound deposits, *in, Proceedings of IAMG Berlin*, September 2002.

Frostick, L. E., and Reid, I., Tectonic Control of Desert Sediments in Rift Basins Ancient and Modern, *in, Frostick, L. and Reid, I., Desert Sediments: Ancient and Modern*, Geological Society of London Special Publication No. 35, 1987, p. 53-68.

Gold Fields Ghana Limited, *A Technical Report of the Tarkwa Gold Mine, Ghana*, Gold Fields Ghana Ltd, and Iamgold Corporation, http://www.iamgold.com/reports/etc/Tarkwa-TechRpt.Final.pdf, 2003, 50p.

Gower, J. C., A General Coefficient of Similarity and Some of Its Properties, *Biometrics,* vol. 27, p. 857-874.

Hirdes, W., and Nunoo, B., The Proterozoic Paleoplacers at Tarkwa Gold Mine, SW Ghana: Sedimentology, Mineralogy, and Precise Age Dating of the Main Reef and West Reef, and Bearing of the Investigations on Source Area Aspects, *in,* Oberthur, T., (ed.), *Metallogenesis of Selected Gold Deposits in Africa*, Geologisches Jahrbuch, Reihe D, Heft 100, 1994, p. 247-311.

Journel, A. G., and Kyriakidis, P. C., *Evaluation of Mineral Reserves: A Simulation Approach*, Oxford University Press, New York, 2004, 216p.

Middleton, G. V., Johannes Walther's Law of the Correlation of Facies, *Geological Society of America Bulletin*, vol. 84, p. 979-988.

Sestini, G., Sedimentology of a Paleoplacer, The Gold-bearing Tarkwaian of Ghana, *in* Amstutz, G. C., and Bernard, A. J., (eds.), *Ores in Sediments*, IUGS, Series A, No. 3, Springer-Verlag, 1976, p. 275-305.

Sylvester, P. J., and Attoh, K., Lithostratigraphy and Composition of 2.1 Ga Greenstone Belts of the West African Craton and Their Bearing on Crustal Evolution and the Archean-Proterozoic Boundary, *Journal of Geology*, vol. 100, p. 337-393.

van Laarhoven, P. J. M., and Aarts, E. H. L., *Simulated Annealing: Theory and Application*, D. Reidel Publishing Company, Dordrecht, 1987, 186p.

Ward, J., Hierarchical Grouping to Optimize an Objective Function, *Journal of the American Statistical Association,* vol. 58, 1963, p. 236-244.

Weissmann, G. S., Carle, S. F., and Fogg, G. E., Three-Dimensional Hydrofacies Modeling Based on Soil Surveys and Transition Probability Geostatistics, *Water Resources Research,* vol. 35, no. 6, 1999, p. 1761-1770.

Wishart, D., k-Means Clustering with Outlier Detection, Mixed Variables, and Missing Values, *Proceedings of GfKl 2001 (25th Annual Conference of the German Classification Society, The University of Munich, 14-16 March, 2001, Munich).*

CONDITIONAL SIMULATION OF GRADE IN A MULTI-ELEMENT MASSIVE SULPHIDE DEPOSIT

N.A. SCHOFIELD
Hellman and Schofield, P.O. Box 599, Beecroft, NSW 2119, Australia

Abstract. In the past decade, conditional simulation methods have been used widely to model the distribution of grades in precious and base metal deposits. Often a number of simulations are used as a basis for evaluating the risk of ore and waste misclassification and improving ore selection practices. The complexity of the application of simulation methods can depend on the nature of the mineralization being modelled. A single element deposit such as gold in a disseminated style of mineralization typical of epithermal gold deposits may be an example of a less complex application. Multi-element deposits with multiple geostatistical sample populations in complex structural settings such as the Cannington silver-lead-zinc deposit represent more challenging applications. This paper discusses a method to generate relatively large scale conditional simulations of mineralization geometry and multiple elements in such deposits.

1 The Mineral Deposit

The Cannington base metal deposit was discovered in 1990 by BHP Minerals (Bailey, 1998). The silver-lead-zinc mineralization is associated with a diverse package of siliceous and mafic rocks with extensive retrogression and alteration. A zoning of base metals is evident within the Southern Zone which is consistent with the interpreted isoclinal fold structure. The lode horizons are defined by the spatial distribution of the base metals. The mineralization types totalling 10 to date, describe the geometry, economic, geochemical and textural relationships within the deposit. Locally, the mineralised sequence around the fold shown in Figure 1 is commonly referred to as the Footwall (CW, NS and CK mineralization types), Hanging-wall (BM, BL and KH) and Hinge (GH, GHB) areas. Mining began in the rich silver-lead-zinc concentrations hosted mainly within the Glenholme mineralization (GH, GHB) within the Hinge area.

2 Modelling using Conditional Simulation

Since early 2000, the Cannington mine geologists have been experimenting with modelling of the distribution of mineralisation types and mineralisation grade using a combination of Probability Field conditional simulation (PF) (Froidevaux, 1992, Srivastava, 1990) and Sequential Gaussian Simulation (SGS) (Gomez-Hernandez and Journel, 1992). The process is ongoing. The focus of this paper is the distribution of

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 449-455.

lead, zinc and silver within a suite of mineralised units in the Hinge area marked GHB in Figure 1.

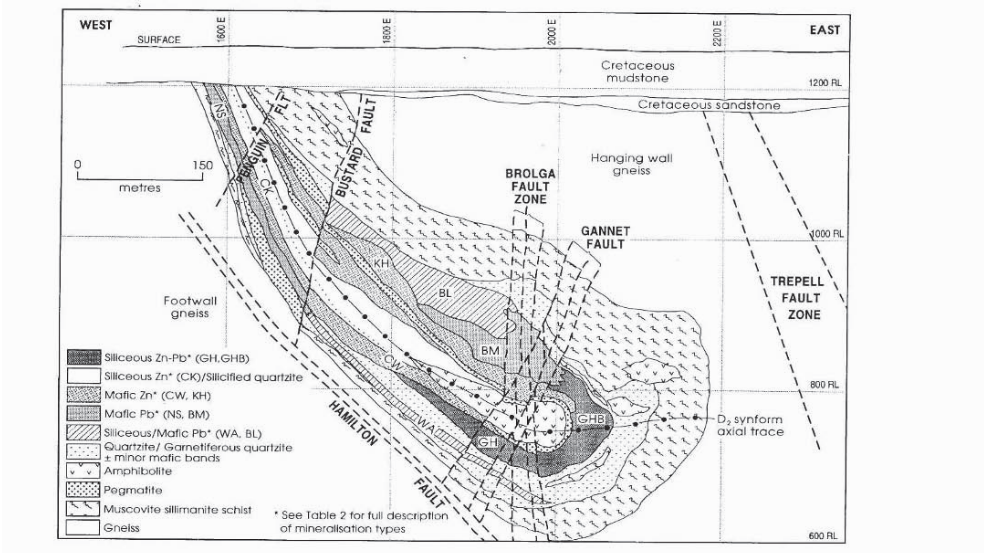


Figure 1: Geologic section through the Cannington Deposit (after Bailey, 1998)

3 The Drill-Hole Data and Statistics

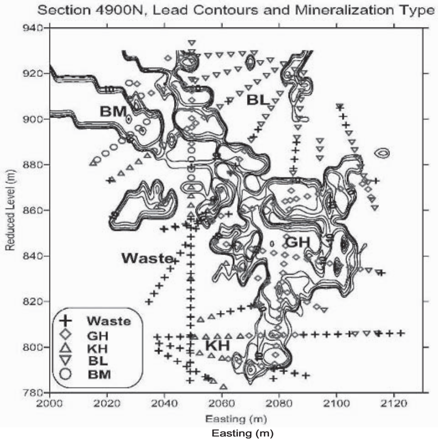


Figure 2: Typical drill-hole cross

Figure 2 presents a typical drill-hole cross-section through the mineralization in the Hinge area. The spatial distribution of the logged mineralisation types is shown by different symbols while the contours map the lead grade concentration in the drill holes. The higher grades of lead and silver occur in the Glenholme (GH) mineralization which forms the central body in the Hinge area. The geometry of the mineralization types varies along the northerly extension of the fold nose, pinching and swelling in response to structural influences. The overall trend of the mineralised structure is around N12E and inclined at around 10 degrees. The lower grade KH mineralization occurs to the west of the GH while the lower grade BL mineralization occurs to the east of and above the GH. The hanging wall lead mineralization (BM) occurs in the upper left quadrant of the section.

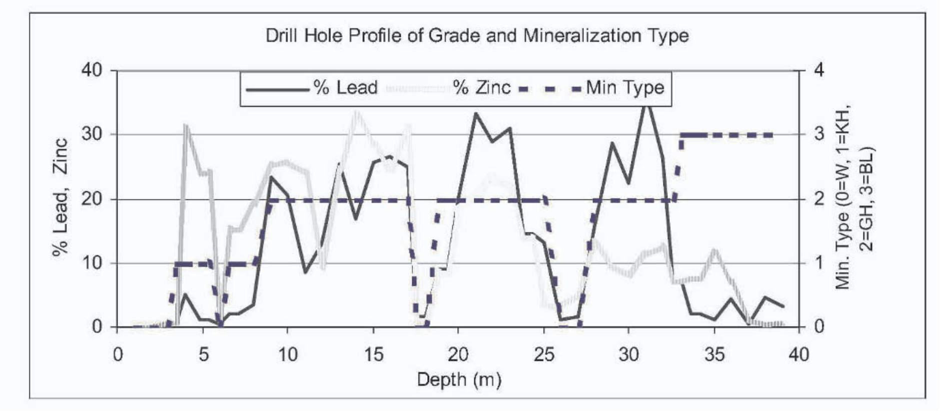


Figure 3: Typical drill-hole profile of lead-zinc and mineralization type.

	Lead %				Silver ppm				Zinc %			
	KH	GH	BL	BM	KH	GH	BL	BM	KH	GH	BL	BM
Mean	2.07	14.28	2.15	8.38	64	573	92	283	4.48	8.58	1.70	5.88
Std. Dev.	4.99	11.83	4.40	10.56	129	563	496	370	6.33	6.08	2.83	5.78
Minimum	0.0	0.0	0.0	0.0	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Median	0.13	11.70	0.28	2.43	22	410	15	96	1.15	7.89	1.15	4.72
Maximum	40.80	60.00	44.25	43.10	1550	10700	21900	2600	33.40	38.50	33.40	35.50
Number	2779	7120	2779	1234	2779	7120	2779	1234	2779	7120	2779	1234

Table 1: Summary statistics of lead, silver and zinc in the mineralization types.

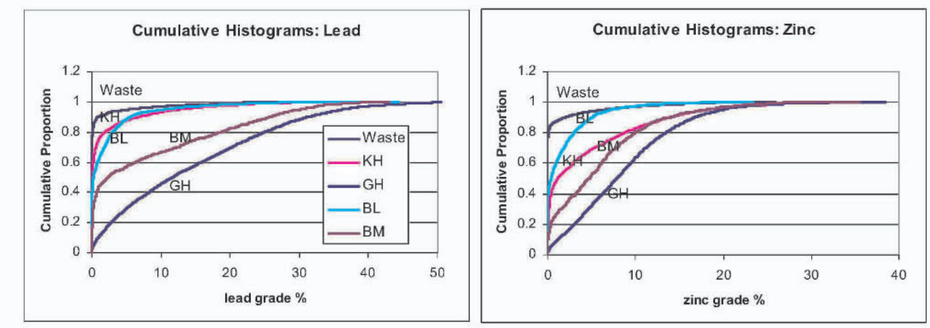


Figure 4: Cumulative histograms of lead and zinc in the mineralization types.

Figure 3 shows a typical drill-hole profile across the mineralization emphasising the sharp changes in lead grade that occur at the transition of one mineralization type to another. Silver tends to follow lead closely but not so zinc as the figure shows.

Summary statistics of lead, silver and zinc in all mineralization types are shown in Table 1. Figure 4 presents cumulative histograms of lead and zinc in all mineralization types. The economic dominance of the GH mineralization is obvious from average concentrations of lead, silver and zinc in this mineralization. Bivariate correlation matrices shown in Figure 5 indicate lead and silver are strongly correlated in all mineralization types. Linear correlations are shown in the upper triangle and the rank order correlations are shown in the lower triangle.

GH Mineralization			
Elements	Lead	Silver	Zinc
Lead	1.000	0.820	0.515
Silver	0.910	1.000	0.296
Zinc	0.662	0.532	1.000
KH Mineralization			
Lead	1.000	0.904	0.516
Silver	0.827	1.000	0.434
Zinc	0.720	0.681	1.000
BM Mineralization			
Lead	1.000	0.901	0.456
Silver	0.942	1.000	0.365
Zinc	0.701	0.672	1.000

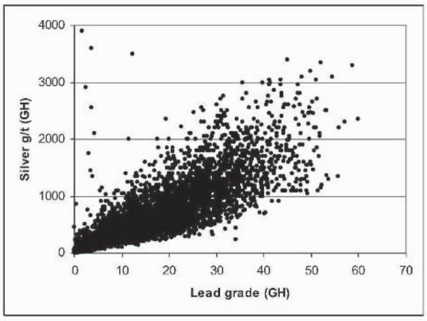


Figure 5: Element correlations and silver-lead scatter-plot in GH.

4 Spatial Continuity of Mineralization Geometry

A numerical sequence is used to describe the mineralization types. Where possible, the sequence should correspond to the physical geological sequence such as the stratigraphic sequence or a nested pattern of alteration. The numerical sequence used in the present case is shown in Table 2 below.

Mineralization	Waste	KH	GH	BL	BM
Sequence no.	0	1	2	3	4

Table 2: Numerical coding of the sequence of mineralization types

The numerical sequence allows continuity of the mineralization types to be described in terms of indicator continuity functions. These functions describe the transition from one set of mineralization types to another e.g. the indicator continuity function for the threshold 0.5 describes the transition from the Waste type to all other mineralization types. Figure 6 shows the geometry indicator continuity maps for the four transition thresholds in the horizontal plane.

5 Spatial Continuity of Lead, Silver and Zinc Grades

The normal scores transforms (Deutsch and Journel, 1992) of these elements are used to describe their spatial continuity within each mineralization type. Inadequate data or the complexity of the mineralization geometry can cause difficulties in the description of these continuities for some mineralization types. In such cases, the element continuity properties of the dominant mineralization type may be adopted or the data from two or more mineralization types may be combined to provide a more appropriate description of the element continuities for those types.

6 Conditional Simulation of Mineralization Types and Grades

The spatial distribution of the mineralization types can be simulated using the method of Probability Field (PF) simulation proposed by Srivastava 1990. In the present case, the non-conditional field of probabilities is generated with the Sequential Gaussian Simulation method (SGS). The conditioning of the probability field to the local mineralization type data is based on the estimate of cumulative probability for each threshold in the numerical sequence of mineralization types shown above. The approach is a very fast way to generate large scale conditional simulations of categorical variables which are both geologically and statistically acceptable. Table 3 below compares the proportions of simulated mineralization types in two conditional simulations compared to the data coding. The differences between the simulations and the data are mainly due to data clustering.

Volume Proportion of Mineralization Types

	Waste	KH	GH	BL	BM
Samples	0.321	0.140	0.358	0.119	0.062
Simulation 1	0.387	0.099	0.306	0.140	0.068
Simulation 2	0.376	0.099	0.305	0.145	0.074

Table 3: Volume proportions of mineralization types in samples and two simulations.

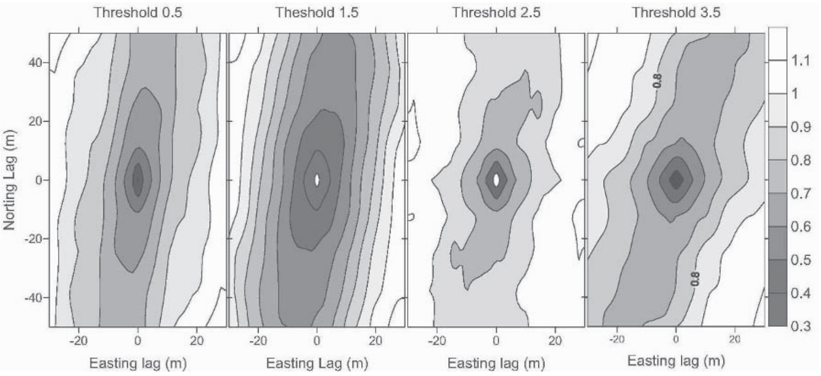


Figure 6: Maps of the sample indicator continuity of mineralization geometry

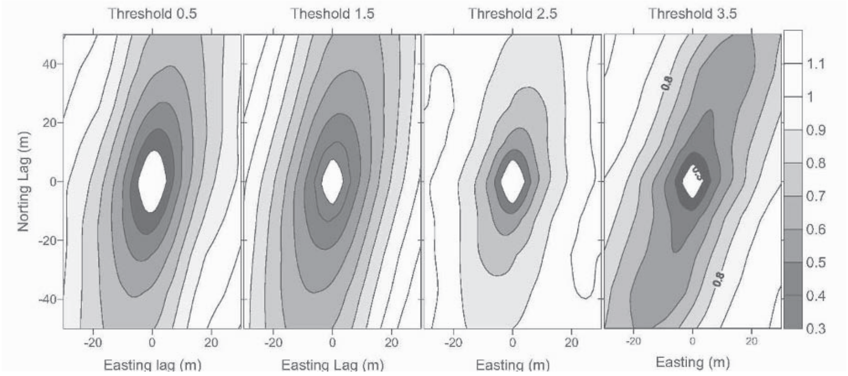


Figure 7: Maps of simulation indicator continuity of mineralization geometry

The continuity maps of the simulated mineralization types shown in Figure 7 show greater continuity at short scale than that of the data shown in Figure 6. This is a deliberate choice to make the maps of mineralization types more realistic, generating relatively smooth contacts between the mineralization types. Simulations of the metal grade distributions for each mineralization type using SGS are found to honour reasonably the univariate, bivariate and spatial properties of the metals (not shown).

Within each mineralization type, the metal distributions are simulated to be independent of the mineralization type boundary as suggested in Figure 3. Figure 8 presents section maps showing the spatial distribution of simulated mineralization types and simulated lead grades.

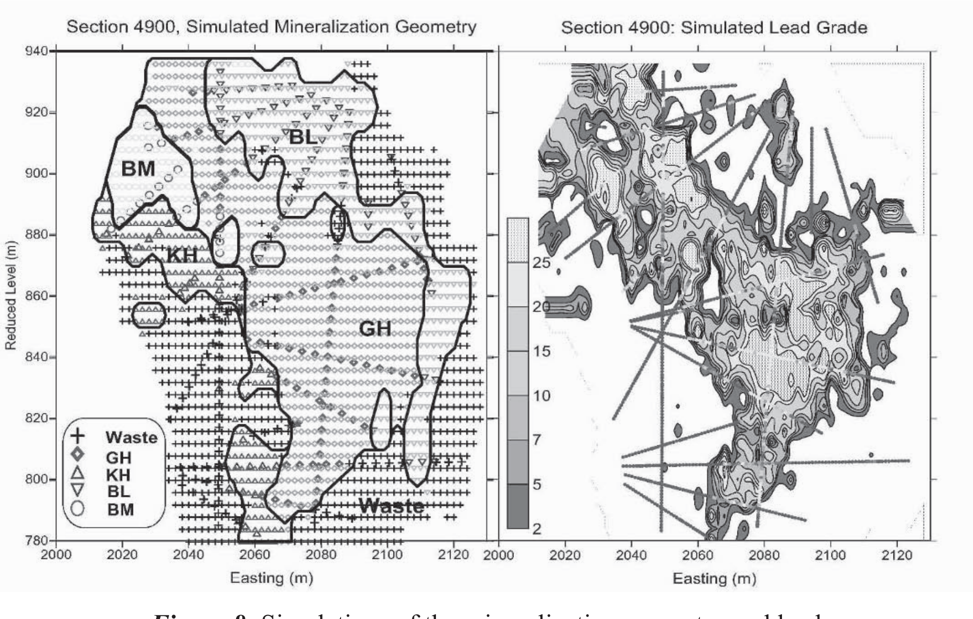


Figure 8: Simulations of the mineralization geometry and lead

7 Conclusions

The Cannington lead-zinc-silver mineralization presents a significant challenge to resource modelling using conditional simulation. The spatial distribution of the mineralization types is complex and dominates the problem of ore definition.

The approach described in this paper generates spatial conditional simulations which reasonably reproduce the statistical and spatial properties of the input data and provide plausible models of the distribution of mineralization types and metal grades.

Results of using the simulations for stope planning and grade prediction indicate the simulations provide an improved basis for locating stopes and reliable predictions of overall stope grades compared to the manual modelling of mineralization type distributions and ordinary kriging estimation of grades. Simulations also provide a better appreciation of the variability in the ore grade that will be realized over time in the mining of stopes.

Acknowledgements

The author acknowledges the cooperation of BHP-Billiton, in particular Alan Edwards in undertaking the work that forms the basis of this paper.

References

Bailey, A., Cannington silver-lead-zinc deposit. in *Geology of Australian and Papua New Guinea Mineral Deposits*, (Eds: A. A. Beckman and D.H. Mackenzie), pp 782-792, 1998. The Australian Institute of Mining and Metallurgy, Melbourne, Australia.

Deutsch, C.V. and Journel, A.G., GSLIB Geostatistical Software Library and User's Guide. Oxford University Press, New York, 1992.

Gomez-Hernandez, J.J. and Journel, A.G., Joint Sequential Simulation of MultiGaussian Fields. In *Geostatistics Troia `92* (Ed. A. Soares), Volume 1, pp85-94, 1992. Kluwer Academic Publishers, London.

Froidevaux, R. Probability Field Simulation. In *Geostatistics Troia `92* (Ed. A. Soares), Volume 1, pp 73-84, 1992. Kluwer Academic Publishers, London.

Srivastava, R.M., An Application of Geostatistical Methods for Risk Analysis in Reservoir Management. 1990. *SPE Paper No 20608.*

This page intentionally blank

THE ULTIMATE TEST – USING PRODUCTION REALITY. A GOLD CASE STUDY

PAUL BLACKNEY, CHRISTINE STANDING and VIVIENNE SNOWDEN
Snowden Mining Industry Consultants
PO Box 77, West Perth, Western Australia, AUSTRALIA, 6872

Abstract. The McKinnons gold mine is owned by Burdekin Resources NL ('Burdekin') and is located within the Cobar Basin in New South Wales, Australia. Open pit mining began in February 1995 and was completed in December 1996. High grade, low grade and mineralised waste stockpiles were processed in separate campaigns up to April 2000 and this production data provides a unique opportunity to test the accuracy of industry standard estimation techniques.

A multi-support data set is available from the mine, comprising exploration reverse circulation (RC) and diamond core (DD) drillholes (25 m sections), RC grade control drilling (12 mE by 12 mN and 6 mE by 6 mN patterns) and some blasthole data. This information was used to create comparative feasibility-stage and grade control resource estimates which were compared with each other and with the actual tonnes and grades processed.

Using the wide-spaced data, feasibility-stage resources were estimated using ordinary kriging, multiple indicator kriging, uniform conditioning, conditional simulation and a global change of support method. Kriging neighbourhood analysis and conditional bias tests were used to determine appropriate panel sizes for estimation. Some models were deliberately constructed using panel sizes that led to conditional bias to test the effect of bias in mine reconciliation. Comparative grade control estimates were then created from the close-spaced data to demonstrate the effects of this additional information.

Estimations were found to be relatively insensitive to estimation method and the band of uncertainty for models based on exploration data was within ± 10% in terms of tonnes and grade. Although confidence limits improved significantly to within ± 3% in tonnes and grade for grade control estimates, it was found that estimates were sensitive to whether hard or soft boundaries were used to control the estimation.

The study illustrates that the bounds of uncertainty reflected by conditional simulations are indicative of the confidence limits for a given choice of geological model and the associated geostatistical parameters and estimation technique. Any changes in these inputs will all have some affect on the accuracy of the estimate. Even production estimates may be affected by perception, particularly related to the effective cut-off applied.

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 457-468.

Practitioners should be aware that resource risk will include cumulative errors arising from a complex process. Once projects are in production it is important to check and validate the key assumptions made during the project evaluation. The only way to achieve this satisfactorily is through the process of reconciliation, that is the comparison of actual production (tonnes, grade and metal) with predictions (resources, reserves and mine plans). This is an often-ignored but vital aspect of the mine value chain, and allows a reality check on the feasibility process which can trigger remedial action if necessary.

1 Introduction

During closure of the McKinnons gold mine, Burdekin compiled a comprehensive data package containing feasibility-stage data, mining grade control data and gold production results for high and low grade ore types. Elliot et al (2001) described the reconciliation of production data compared with the feasibility study estimates prepared before mining began in 1995. The input data has now been used in a study to quantify the accuracy of current day industry standard estimation techniques used to estimate resources in the Australian gold mining sector.

2 Outline of study

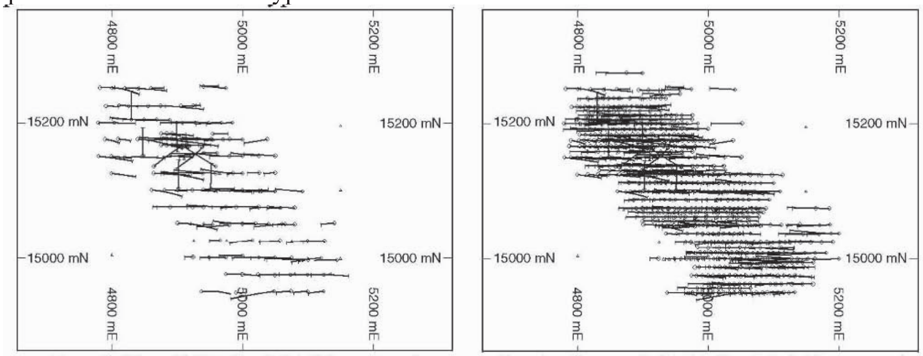
There are a number of nodes of uncertainty which can impact on resource estimation, including data integrity, geological interpretation, grade estimation error and ore mining control. The uncertainty may be demonstrated by comparing the tonnes, grade and metal estimates of planning models with production reality. This is currently topical in Australia because of the proposed revision to the JORC Code (JORC 1999) which encourages the Competent Person to quantify risk/uncertainty associated with resource/reserve estimates.

In this study, resource models were created using the wide-spaced drillhole and mineralisation domains as interpreted during the McKinnons feasibility study. Comparative grade control models were created from the close-spaced mining data, more detailed mineralisation domains and, for some estimates, the production grade control polygons as mined. Estimation methods included ordinary kriging (OK), multiple indicator kriging (MIK), sequential indicator simulation (SISIM), uniform conditioning (UC) and global change of support (GCOS) methods. Model block sizes were determined from kriging neighbourhood analysis and conditional bias tests. The actual tonnages and grades processed during the life of the mine were used to test the accuracy of the exploration and grade control models using the various estimation methods.

3 Background

Between 1990 and 1994, RC and diamond drilling on 25 m sections was used to define the resource for feasibility studies. Following the decision to mine in 1994, the entire deposit was drilled on a 12 mE by 12 mN pattern with some infill on a 6 mE by 6 mN pattern (*Figure 1*) to provide pre-production grade control modelling.

All high grade ore (>1.3 g/t Au) was processed by November 1997 and low grade stockpiles (0.7 to 1.3 g/t Au) were processed by October 1998. Processing of mineralised waste (0.3 to 0.7 g/t Au) was completed in April 2000. The actual production for each ore type is listed in *Table 1*



Feasibility (213 dh, 10,701 comps) Grade Control (547 dh, 35,649 comps)

Figure 1 McKinnons drillhole location trace plans.

Ore	K tonnes	Grade
High Grade (>1.3 g/t)	1,204	2.64 g/t
Low Grade (0.7 to 1.3 g/t)	399	1.08 g/t
Mineralised Waste (0.3 to 0.7 g/t)	1,067	0.65 g/t
Total	**2,670**	**1.61 g/t**

Table 1 McKinnons production data (Elliot et al 2001).

4 Estimation methods

Kriging neighbourhood analysis and conditional bias tests were employed prior to estimation to determine a suitable block size for unbiased estimation (Krige, 1996). OK is optimal for near normal distributions (Goovaerts, 1997) but is often used in estimation of skewed data, together with capping or top cutting of sample grades. MIK and UC are non-linear techniques which accommodate highly-skewed or mixed data distributions (Glacken and Snowden, 2001) and deliver recoverable resources which represent the appropriate selective mining unit (SMU).

SISIM (Deutsch and Journel, 1998) can be used to produce a number of equally-probable realisations which can be re-blocked to represent the dimensions of a target SMU and reported to determine the probability and grade above cut-off for each SMU. The tonnage and grade spread of the simulation results reflect the confidence limits in tonnage and grade and can be used to derive the confidence limits associated with a given mining area or period.

5 Feasibility data analysis and modelling

For this study, the available datasets and subsequent models were clipped to the limits of the end-of-mine open pit survey to define a consistent space for comparison. The statistical and spatial character of the feasibility data was investigated within subdomains of a global envelope defined at a grade threshold of ~0.1 g/t Au. Within this envelope three main statistical domains were defined, comprising a central high grade domain separating two lower grade domains. The geology is complex and mineralisation is believed to be structurally controlled and so domains were essentially defined by grade. The gold grade distributions within all domains displayed a high positive skewness, with coefficients of variation of approximately 2.5.

Gold grade continuity was primarily investigated using standardised indicator variograms. The indicator analysis revealed nugget effects representing 10% to 43% of the total sill, with most thresholds displaying a nugget effect of 20% to 35% of the total sill. All three domains displayed patterns of rotating anisotropy associated with increasing grade. The variogram model parameters from these analyses were used to control MIK estimation and SISIM computations. OK estimation was based on back-transformed variogram model parameters.

A kriging neighbourhood analysis, using the median indicator variogram, was employed to determine the optimal kriging block size for the high grade central domain (Krige, 1996). From a range of block sizes and locations tested, a 12.5 mE by 12.5 mN block on a 2.5 m bench was identified as optimal, with this block size returning regression slopes of up to 0.92 and kriging efficiencies of up to 68%. OK and MIK models were estimated using this block size and conditional bias sensitivity models were created using block sizes of 5 mE by 5 mN on 2.5 m benches and 25 mE by 25 mN on 5 m benches.

Simulation models were based on 100 realisations generated on a node spacing of 2.5 mE by 2.5 mN by 1.25 mRL. The realisations were re-blocked to supports of 5 mE by 5 mN and 12.5 mE by 12.5 mN blocks on 2.5 m benches, and 25 mE by 25 mN blocks on 5 m benches, and presented in terms of the median and 90% confidence limits.

6 Grade control data analysis and modelling

The grade control (GC) data comprised all the original wide-spaced exploration drillholes plus the additional close-spaced RC drilling completed for grade control (*Figure 1*). Categorical indicator kriging using a 0.1 g/t Au indicator grade was applied within three large scale domains to define the grade control mineralised envelope. The categorical model was visually compared to the input drillhole data to determine a probability threshold that closely reproduced the spatial patterns evident in the drillholes (*Figure 2*, left). The mineralised envelope was then further domained into seven regions that reflected a weakly developed depletion zone overprinting primary grade continuity trends (*Figure 2*, right).

The domaining was successful in producing singular distributions for the lower grade domains but mixed distributions were still evident in the higher grade domains. An indicator method was therefore selected for grade estimation. The indicator variography analysis revealed nugget effects ranging from 40% to 75% of the total sill and ranges of a few metres to 30 m vertically and from 15 m to 190 m in the horizontal plane depending on domain and indicator grade. Two of the seven domains exhibited rotating anisotropy associated with increasing indicator grade.

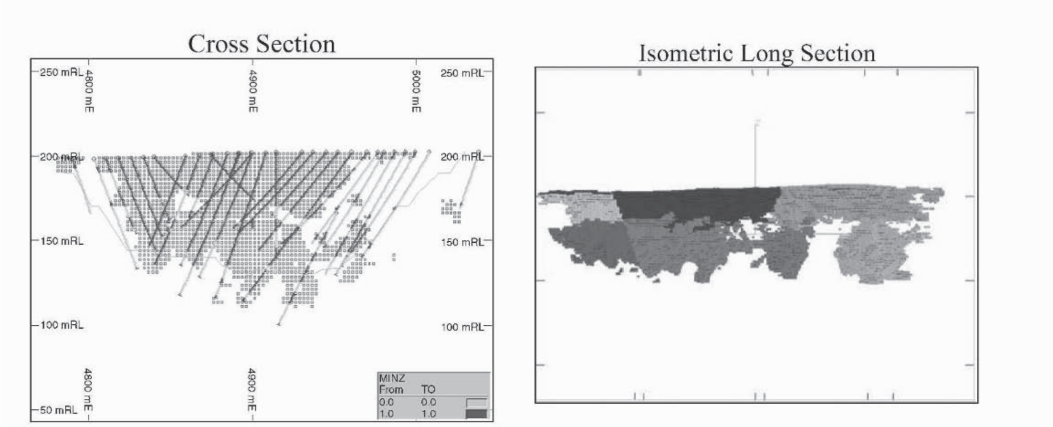
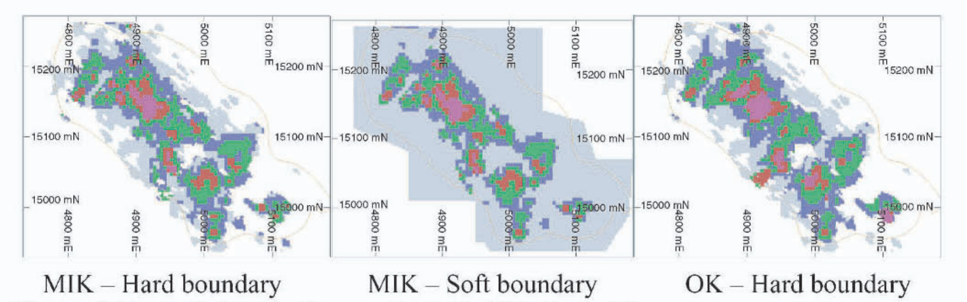


Figure 2 GC sections, 0.1 g/t Au indicator section (left) and structural domains (right).

Block grades were estimated using both MIK and OK. The grade control estimation block size was set to 5 mE by 5 mN on 2.5 mRL benches to reflect the expected mining selectivity. Two MIK estimates were computed, one with a hard boundary (HB) between structural domains and a second using a soft boundary (SB) condition. The OK estimate was computed using the hard boundary and grade top-cuts, which varied by domain. *Figure 3* illustrates the presentation of these three models on a typical bench.



MIK – Hard boundary MIK – Soft boundary OK – Hard boundary

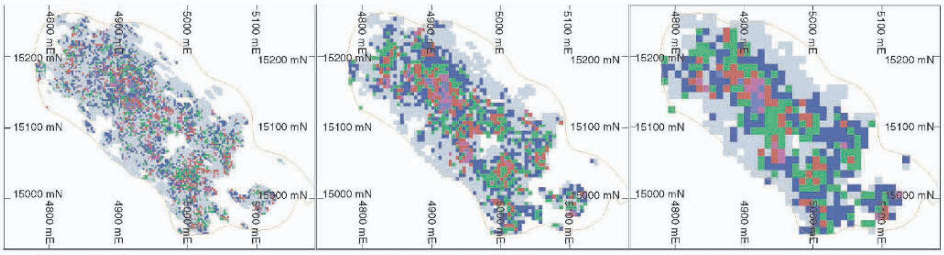
Figure 3 Example bench plans grade control block models.

A kriging neighbourhood analysis using the OK variogram later revealed significant variation in regression slope and kriging efficiency by domain. For example, in a well drilled domain, 96% of the model blocks have regression slopes better than 0.9 and 83% of the blocks have kriging efficiencies better than 70%. In contrast, in a more poorly drilled domain, only 66% of the blocks have regression slopes that are better than 0.9 and only 21% of the blocks have kriging efficiencies better than 70%.

One hundred realisations of the grade control data were generated using SISIM. The domain controls and parameters for the simulation were the same as those applied to the HB MIK. A node spacing of 2.5 mE by 2.5 mN by 1.25 mRL was employed

Simulations were re-blocked to represent the behaviour of 5 mE by 5 mN by 2.5 mRL and 10 mE by 10 mN by 2.5 mRL SMU's for comparison with the kriged models (*Figure 4*).

Polygons representing the production grade control interpretation were available for the upper part of the pit. This interpretation was used to create a traditional polygonal model by calculating the top-cut average grade of the grade control samples located within each polygon.



Simulation nodes SMU (5x5x2.5) SMU (10x10x2.5)

Figure 4 Example median grade control SISIM bench plans

7 Comparative Results

7.1 EXPLORATION MODELS

7.1.1 OK versus MIK

The OK and MIK exploration models demonstrated very little global sensitivity to estimation method. Estimation selectivity showed only marginal changes as the block size was increased from 5 mE by 5 mN by 2.5 mRL to 25 mE by 25 mN by 5 mRL due to the diffuse nature of the mineralisation and the relatively high nugget effect.

7.1.2 Kriging versus UC versus GCOS

The closest comparison between the theoretical GCOS and UC estimates of the 6.25 mE by 6.25 mN by 2.5 mRL SMU grade-tonnage relationship occurred for the 25 mE by 25 mN by 5 mRL OK model. This is despite the kriging neighbourhood analysis supporting the smaller 12.5 mE by 12.5 mN by 2.5 mRL panel size for kriging. This result probably reflects the limitations of the simple kriging neighbourhood analysis completed for this study and demonstrates that overall, estimation into the larger panel size suffers less conditional bias.

7.1.3 Kriging versus simulation

The comparisons between the exploration OK, MIK and SISIM models are illustrated in *Figure 5*. The models are based on a block size of 12.5 mE by 12.5 mN by 2.5 mRL. There is a bias noted whereby the average grades of the OK and MIK models are located towards the upper confidence limits defined by the SISIM models which is believed to be due to differences in treatment of the high grade tail of the grade distribution.

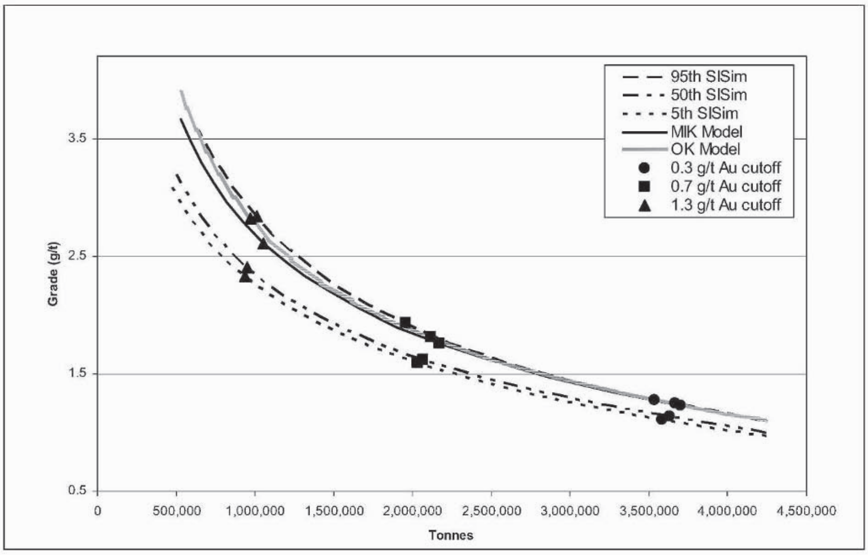


Figure 5 Grade-tonnage curves of the exploration OK, MIK and SISIM models.

The simulations validate the optimal model block size determined by kriging neighbourhood analysis as the kriged block models reflect a similar grade/tonnage profile to the simulations. The OK model appears to be slightly oversmoothed compared with the MIK model.

Using SISIM to quantify confidence limits, and making the assumption that the range of simulations between the 5th and 95th percentiles give a reasonable representation of the space of uncertainty, the 90% confidence limit at a cut-off of 0.3 g/t is ±1% on tonnage and ±7% on grade. At a 0.7 g/t cut-off, the confidence limit is ±2% on tonnage and ±10% on grade. The tonnage uncertainty again increases at a 1.3 g/t cut-off, where the confidence limit is ±4% on tonnage but the confidence limit remains at ±10% in grade.

7.2 GRADE CONTROL MODELS

7.2.1 OK versus HB MIK versus SB MIK

Grade-tonnage reporting from the HB MIK, SB MIK and HB OK models within pit is presented in *Figure 6*. The highest metal profile is presented by the HB OK model, followed by the HB MIK and then the SB MIK models. Tonnage and grade reporting is similar between all estimates at the 0.7 g/t and 1.3 g/t cut-offs, however the influence of the soft boundary assumption is readily apparent at the 0.3 g/t cut-off. Tonnage and grade predictions at the 0.7 g/t and 1.3 g/t cut-offs are within 0% to 8% of each other for all models. At the 0.3 g/t cut-off, the SB MIK model predicts 20% more tonnage at 16% less grade.

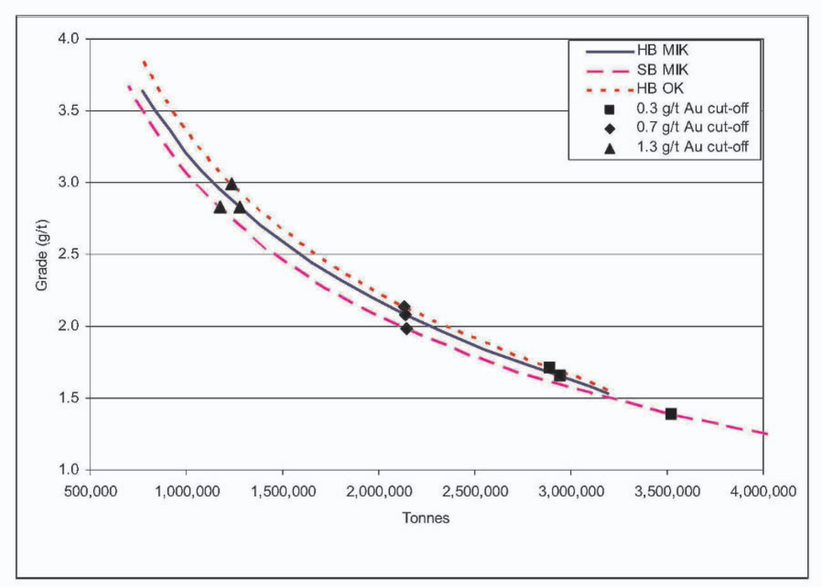


Figure 6 Grade-tonnage comparisons between HB MIK, SB MIK and HB OK grade control estimates.

7.2.2 HB MIK versus SISIM

The comparison of the HB MIK estimate to the SMU predictions provided by SISIM indicated that the grade-tonnage profile of the kriged estimate was closer to that predicted for a 10 mE by 10 mN by 2.5 mRL SMU rather than the 5 mE by 5 mN by 2.5 mRL block size used during estimation (*Figure 7*).

This outcome is somewhat at odds with the results of the kriging neighbourhood analysis which suggested there was minimal conditional bias at a 5 mE by 5 mN by 2.5 mRL block size within the majority of the pit. Some of the mineralisation domains located deeper within the pit may be oversmoothed but it is surprising that the simulations show that the effective resolution of the kriging overall is 10 mE by 10 mN by 2.5 mRL. This outcome may be a function of the relatively high nugget effect shown by the variograms.

Using the 10 mE by 10 mN by 2.5 mRL re-blocked simulations to define the limits of uncertainty for grade control, the 90% confidence limit is ±0% for tonnage and ±3% for grade at a cut-off of 0.3 g/t. At a 0.7 g/t cut-off, the confidence limits are ±1% for tonnage and ±2% for grade. The tonnage uncertainty increases slightly to ±3% at a 1.3 g/t cut-off, but remains at ±2% for grade.

8 Reconciliation of models with production

Actual production is shown, together with the comparable exploration UC and grade control MIK estimates in *Figure 8*. There is minimal uncertainty due to estimation block size, with the UC change of support from the 12 mE by 12 mN by 2.5 mRL model being marginally lower grade than from the 25 mE by 25 mN by 5 mRL model.

The HB MIK grade control model is at the upper limit of the estimates while the SB MIK grade control model presents a grade-tonnage profile closer to that of the production reporting, except that this model suggests the presence of considerably more tonnage than realised at a cut-off of 0.3 g/t. If the HB MIK model is considered to be appropriate, there may have been potential to improve the metal recovery (*Figure 9*).

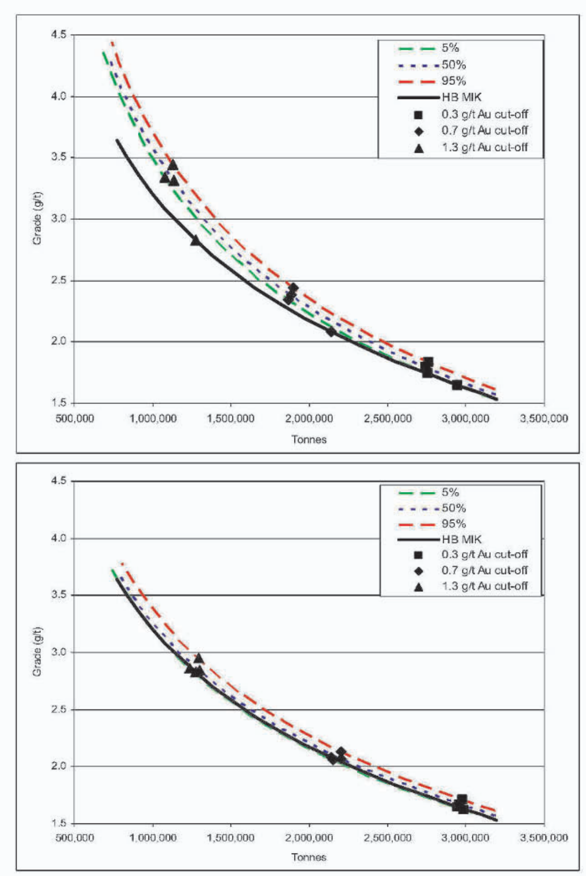


Figure 7 Grade-tonnage of HB MIK (5 mE by 5 mN by 2.5 mRL) with 90% confidence limits of SISIM at 5 mE by 5 mN by 2.5 mRL support (top) and 10 mE by 10 mN by 2.5 mRL support (bottom).

At the high grade cut-off of 1.3 g/t, the HB MIK model suggests slightly more tonnes at a higher grade might have been achieved. At the cut-off of 0.7 g/t, the model suggests considerably more tonnage could have been recovered at a similar head grade and at the 0.3 g/t cut-off more tonnage could have been recovered.

The 0.7 g/t production estimate represents less tonnes at higher grade than any of the models, perhaps suggesting the actual SMU cut-off applied during mining was higher than 0.7 g/t. This could well be due to the polygonal grade control approach applied to unsmoothed sample grades and would be in line with the volume-variance relationship.

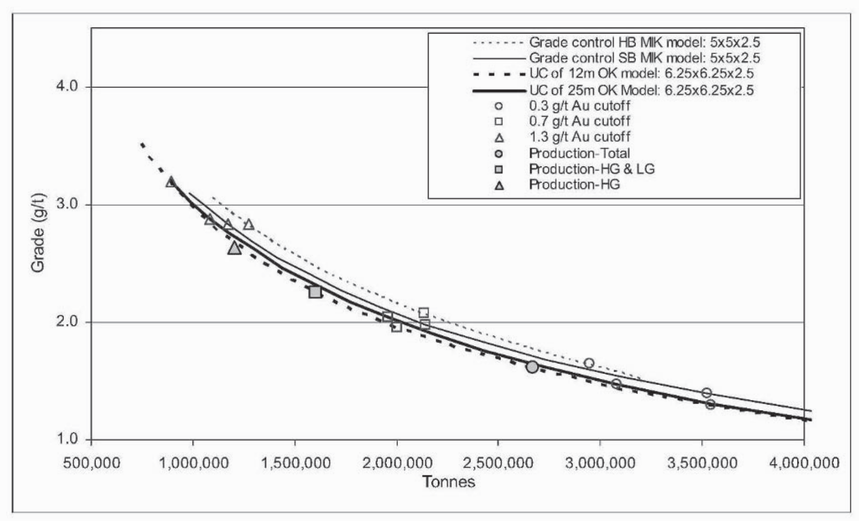


Figure 8 Grade-tonnage curves for the exploration and grade control models compared with production results.

The average grades of the HB MIK and SB MIK models within the production outlines as dug agree with those determined by the polygonal modelling (*Figure 10*). However, it is possible some metal was misclassified and the effective lower cut-off of the production model is about 0.5 g/t (as per the block cut-off for both the hard and soft boundary models) rather than 0.3 g/t as reported by production records.

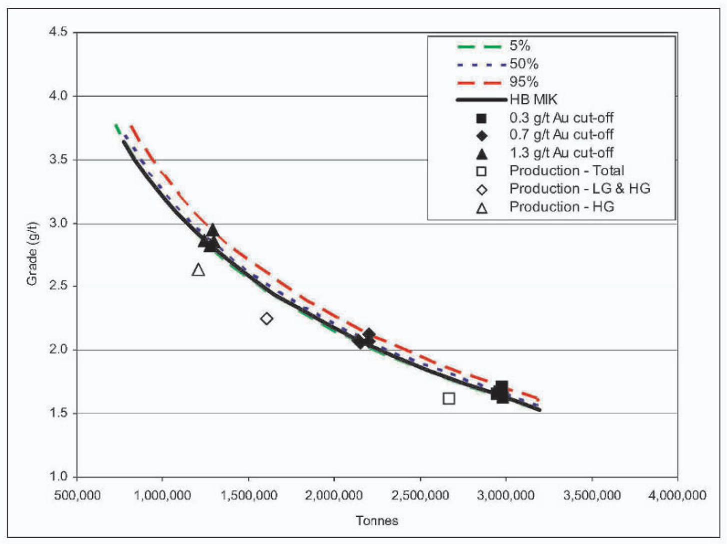


Figure 9 Actual production results and grade-tonnage curves for HB MIK (5 mE by 5 mN by 2.5 mRL) with 90% confidence limits of SISIM at 10 mE by 10 mN by 2.5 mRL support.

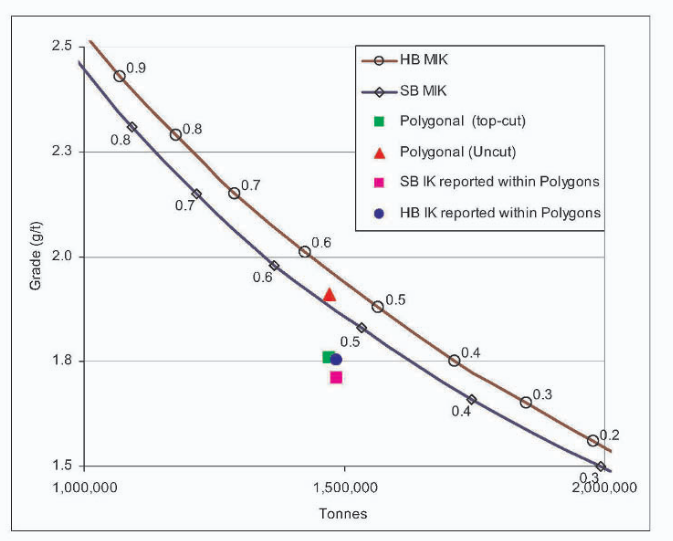


Figure 10 Grade-tonnage reporting above 175 mRL within production interpretation compared to HB MIK and SB MIK.

9 Conclusions

Exploration models show very little sensitivity to estimation method and change of support models using UC are comparable with theoretical GCOS results. The quality of the estimate using 25 mE by 25 mN by 5 mRL blocks with change of support to 6.25 mE by 6.25 mN by 2.5 mRL SMUs is marginally improved compared with estimating into the optimal smaller blocks identified by kriging neighbourhood analysis.

Global 90% confidence limits for 12.5 mE by 12.5 mN by 2.5 mRL blocks based on SISIM of exploration data show tonnage can be estimated within ±1% and grade within ±7% at a 0.3 g/t cut-off. At 0.7 g/t the tonnage uncertainty increases to ±2% and grade to ±10%. At 1.3 g/t, the tonnage uncertainty increases to ±4% and grade uncertainty does not change.

A surprising degree of sensitivity is evident depending on whether grade control models are based on hard or soft domain boundaries and, indeed, whether based on a polygonal estimation method.

Global 90% confidence limits for 10 mE by 10 mN by 2.5 mRL blocks based on SISIM of grade control data show tonnage can theoretically be estimated within ±1% and grade within ±3% at cut-offs of 0.3 g/t and 0.7 g/t cut-off. At 1.3 g/t, the tonnage uncertainty increases to ±3%.

Optimal block sizes defined for exploration models by kriging neighbourhood analysis are confirmed by re-blocking of conditional simulations. However, the effective resolution of the grade control kriging is 10 mE by 10 mN by 2.5 mRL, which is somewhat at odds with the results of the kriging neighbourhood analysis which supports blocks of 5 mE by 5 mN by 2.5 mRL.

The authors have observed complementary behaviours, for example the close alignment of the volume-variance relationship using different techniques, but have also noted unexpected biases, for example at the grade control stage, where the uncertainty should be minimised, yet there remains a sensitivity related to the treatment of domain boundaries.

The learning from this study is that no one approach can be guaranteed fool-proof and, although theoretical tests are partially successful, there will always remain a degree of uncertainty due to inherent variability, geological interpretation and boundary control, estimation technique, scale of mining and human perception. The use of contingencies during feasibility assessments is recommended to determine the impact of both high and low scenarios on the value of the project under consideration.

Acknowledgements

We would like to acknowledge the considerable efforts of M. Franks, S. Hackett and I. Glacken in assisting with this study, Burdekin Resources N.L. for permitting us to use the data from the McKinnons gold deposit for this paper and Snowden Mining Industry Consultants Pty Ltd for providing time and resources for this study to be undertaken.

References

Deutsch, C.V., and Journel, A.G., 1998. GSLIB Geostatistical Software Library and User's Guide. Second Edition. (Oxford University Press: New York).

Elliott, S.M., Snowden, D.V., Bywater, A., Standing, C.A. and Ryba, A., Reconciliation of the McKinnons Gold Deposit, Cobar, New South Wales, *in Mineral Resource and Ore Reserve Estimation – The AusIMM Guide to Good Practice* (Ed: A.C. Edwards), 2001, p. 257-268.

Glacken, I M. and Snowden, D V, Mineral Resource Estimation in Mineral Resource and Ore Reserve Estimation – The AusIMM Guide to Good Practice (Ed: A.C. Edwards), 2001, p. 189-197.

Goovaerts, P., Geostatistics for Natural Resources Evaluation. Oxford University Press. 1997, 483 pp

Joint Committee of the Australasian Institute of Mining and Metallurgy, Australasian Institute of Geoscientists, and Minerals Council of Australia, 1999. Australasian Code for reporting of identified Mineral Resources and Ore Reserves, 1999 edition.

Khosrowshahi, S, and Shaw, W J 1997. Conditional simulation for resource characterisation and grade control – principles and practice, in Proceedings World Gold '97 conference, pp 275-282 (The AusIMM, Singapore, 1997).

Krige, D G, 1996. A practical analysis of the effects of spatial structure and of data available and accessed, on conditional biases in ordinary kriging, in Geostatistics Wollongong '96 (Eds: Baafi E Y, and Schofield, N A) pp 799-810 (Kluwer, The Netherlands, 1997).

ORE-THICKNESS AND NICKEL GRADE RESOURCE CONFIDENCE AT THE KONIAMBO NICKEL LATERITE (A CONDITIONAL SIMULATION VOYAGE OF DISCOVERY)

MARK MURPHY[1], HARRY PARKER[2], ANDREW ROSS[1] and MARC-ANTOINE AUDET[3]
[1]Snowden Mining Industry Consultants, [2]AMEC Americas Limited,
[3]Falconbridge Nouvelle Caledonie SAS

Abstract. Tropical weathering on the ridges of the Koniambo massif in New Caledonia has produced nickel mineralisation of variable thickness. Conditional simulation studies of nickel grade and ore-thickness (a proxy for ore tonnage) were used to quantify the resource risk and to generate constraining envelopes for resource classification.

Ore-thickness intercepts were created from vertical drilling and converted to 2D point data. Many drillholes that did not meet the selection criteria were included as barren, and these holes imposed a strong positive skewness on the data histograms. Indicator variography revealed that both grade and ore-thickness continuity is quasi-isotropic. One-hundred 2D sequential indicator conditional simulations were generated for each attribute on a 10 m by 10 m grid for the three deposit areas. This paper focuses on results from one area, the Centre sector.

The 2D simulation realisations were reblocked to generate a panel mean for each simulation, and the distribution of the 100 panel means were found to be near normal. Tonnages were computed for each panel from the mean simulation thickness and deposit-average bulk density. Relative 90% confidence limits were then computed for each panel using normal distribution assumptions, the panel distribution standard deviations, and the panel means. However, because confidence limits also depend on production rate, the panel relative 90% confidence limits were scaled to the production increment (quarterly, annual) of interest by incorporating assumptions from the standard error of the mean. The rescaled values revealed the risk on an annual production basis was low for both attributes. On a quarterly production basis, the nickel grade risk was low in all areas, but only areas of close-spaced drilling achieved target levels of tonnage risk. The relative 90% confidence risk maps were then used as a guide for resource classification of the deposit and also to focus an infill drilling programme to support a feasibility study to be used by the sponsors to finance the project.

The simulation method takes the local variability of the data into account, a clear advance over traditional estimation variance techniques. The requisite drillhole spacing required to achieve a desired level of confidence varies within the Koniambo deposit, being tighter in high-variability areas (mixture of ore and pinnacles of waste rock) and broader in low variability (more homogenous) areas.

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 469-478

1 Introduction

Falconbridge Nouvelle Caledonie SAS (Falconbridge) required a risk assessment of the resources at the Koniambo nickel laterite project in New Caledonia. The main aim of the assessment was to determine if additional drilling was required to achieve acceptable levels of confidence in ore grade and tonnage for the project feasibility study. The levels of confidence were to be established through 2D conditional simulations of ore-thickness (a proxy for tonnage) and nickel grade.

Arik (1999) and Yamamoto (2001) have considered that confidence measures should reflect the local data configuration and the local variability of data. Their approach is to calculate these components separately. Conditional simulation avoids this, and allows tractable assessment of risk at any anticipated production scale from a single set of data.

2 Geology and mineralisation

The nickel deposits on the ridges and elevated plateaus of the Koniambo massif are typical of the laterites that have developed under tropical weathering conditions on ultramafic bodies throughout New Caledonia (Figure 1, left). Figure 1 (right) is a generalised Koniambo weathering profile that consists of variable thicknesses of limonite and saprolite that reach a combined maximum thickness of approximately 40 m. High-grade ores are characterised by boxworks of garnierite in the saprolite-limonite transition and the upper levels of the saprolite.

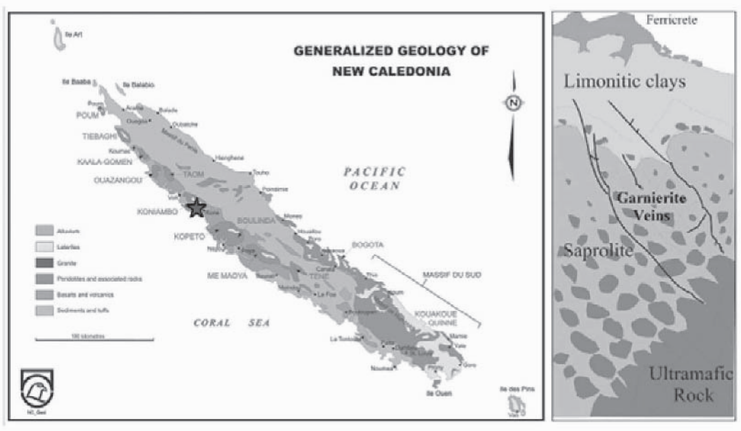


Figure 1 New Caledonia geology (left) and generalised laterite profile (right)

The structural controls and continuity of high-grade nickel mineralisation are not readily identified from vertical drilling. Close-spaced drilling is required to determine the variability of the bedrock topography and the distribution of low-grade boulders and waste pinnacles. Drilling on a 56 m pattern (80 m ‘ quincunx’ pattern) is considered adequate for global resource estimates in the limonitic horizon. However a 28 m pattern (40 m ‘ quincunx’ pattern) is considered necessary to provide confidence in geological interpretations in the more complex saprolite horizons.

3 Input data

The levels of confidence in ore-thickness and nickel grade at Koniambo were evaluated from 2D data because the mineralisation has a very large lateral extent (10 km x 4 km) relative to the depth of the deposit profile.

The study was conducted in two phases. Preliminary runs were made in 2002, and confidence limits obtained indicated the need for 30,000 m of further drilling. The work was updated after the drilling was completed in 2003, and that work is described herein.

Falconbridge divided the deposit into three sectors. This paper focuses on results from the largest area, the Centre sector. The 2D data for both attributes were created from drilling intercepts that met minimum grade and ore-thickness criteria determined by Falconbridge. Figure 2 below shows example locations of the ore thickness data used in the risk assessment study at the Centre sector. The study was constrained to a boundary interpreted to be the limit of mineralisation in the sector.

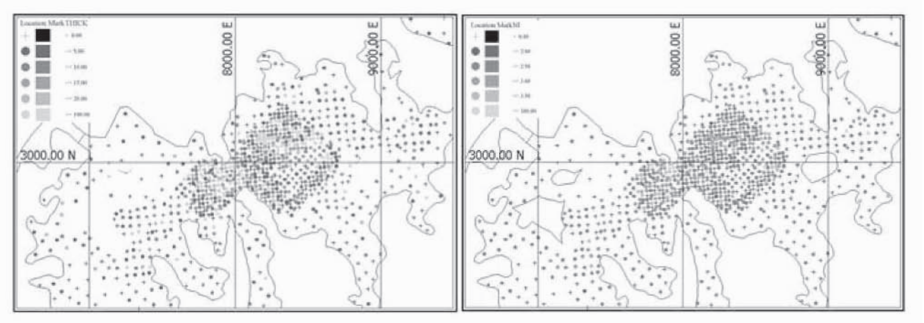


Figure 2 Centre sector 2D data, ore-thickness (left) and nickel grade (right)

Statistics for each attribute were computed using an 80 mE by 80 mN declustering window to account for the variable spacing of drilling (Figure 3). The summary statistics reveal that both attributes have skewed distributions, with over 35% of the data being less than Falconbridge's minimum grade criterion of 2% Ni.

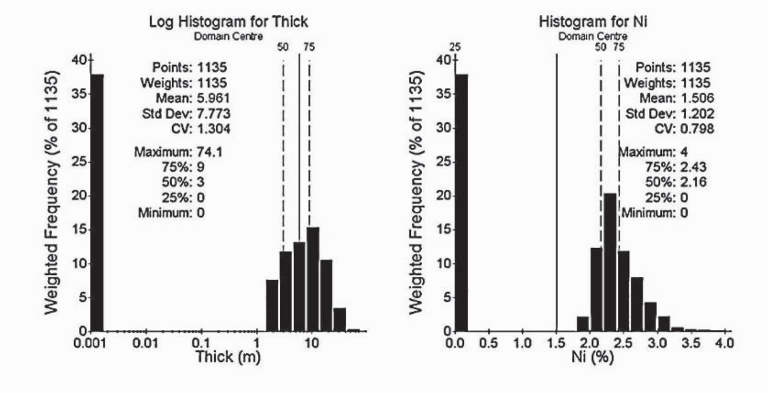


Figure 3 Ore intercept declustered histograms, ore-thickness (left) and nickel grade (right)

4 Variography

Indicator semivariograms were computed for multiple indicator thresholds of the two attributes. The ore-thickness indicator semivariograms are well structured up to the 10 m ore-thickness threshold, although the experimental structure is poor above this threshold. The nickel grade indicator semivariograms reveal a very short-range structure, with the longer range structure declining for the higher thresholds.

Figure 4 and Figure 5 summarise the sill and range values interpreted to fit the experimental indicator semivariograms of ore-thickness and nickel grade of the Centre data. Both attributes show a pattern of decreasing range with increasing indicator threshold. Indicator nugget effects of nickel grade increase with threshold. However the indicator nugget effects of ore thickness are consistent. There is also a pattern of moderate anisotropy for the lower thresholds, but higher thresholds are isotropic.

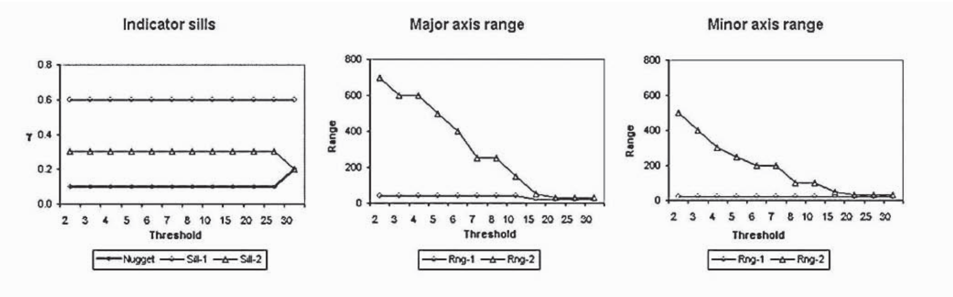


Figure 4 Ore-thickness standardised indicator variography sills (left) and ranges (right)

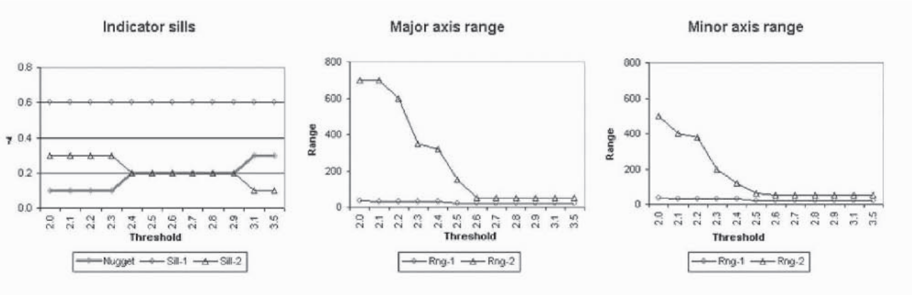


***Figure* 5** Nickel grade standardised indicator variography sills (left) and ranges (right)

The complex changes in variography with changes in threshold also support the choice of multiple indicator simulation instead of sequential gaussian simulation, which would have been an easier method to implement.

5 Simulation

Due to the low-grade spike of zero values associated with each attribute, sequential indicator simulation was selected to provide point-support realisations of ore-thickness and nickel grade. Independent simulation of each attribute was justified because nickel grade is independent of ore-thickness where ore-thickness is greater than zero. Twelve indicator thresholds (as listed on the x-axes Figure 4 and Figure 5) were selected for sequential indicator simulation of the ore-thickness and nickel grade. The first threshold for each attribute was set to partition the barren data, and higher thresholds were set at key attribute values of interest.

Using the input data and indicator variography models, 100 conditional simulations were computed for ore-thickness and nickel grade using the GSLIB, SISIM program for sequential indicator simulation (Deutsch and J ournel, 1998). The simulations were validated by comparing the input statistics and variography with the simulation outputs.

The simulations reproduced the input means of each attribute (Figure 6 and Figure 7), although with marginally lower E-type averages for ore-thickness.

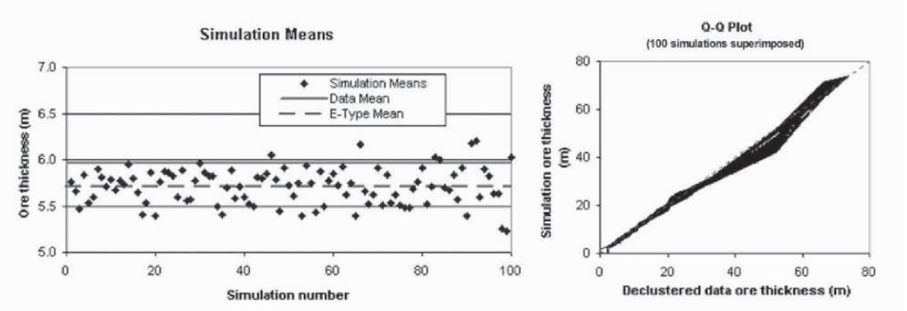


Figure 6 Ore-thickness simulation means (left) and Q-Q plot (right)

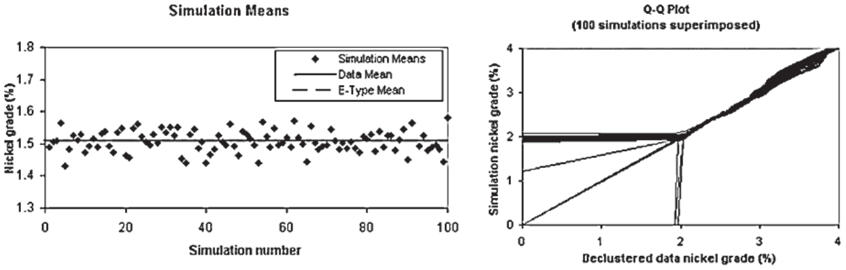


Figure 7 Nickel grade simulation means (left) and Q-Q plot (right)

The data histogram and variogram reproduction was also acceptable for both attributes. However, the E-type estimate of ore-thickness was marginally lower than that of the input data, but the two means were coincident for nickel grade. Images of one realistisation and the E-type averages for ore-thickness and nickel grade are shown in Figure 8 and Figure 9.

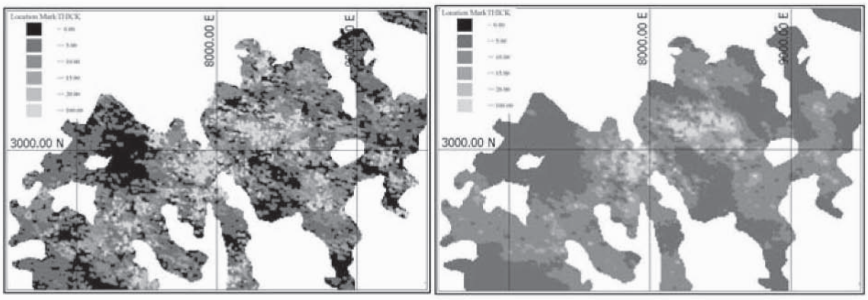


Figure 8 Ore-thickness simulation 001 (left) and E-type average (right)

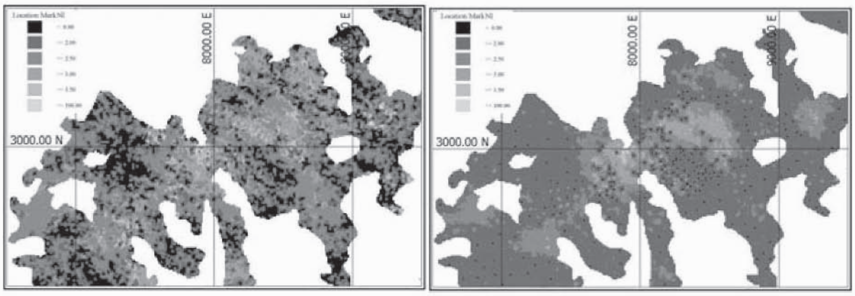


Figure 9 Nickel grade simulation 001 (left) and E-type average (right)

The simulation and E-type images show some spatial correspondence of zones of thicker ore and higher nickel grade. The ore-thickness simulation shows higher relative variability than the nickel grade results.

6 Reblocking and confidence limits

The simulation results of each attribute were averaged or reblocked into 100 m square panels (nominally 100, 10 m by 10 m spaced nodes) to derive a mean value for each panel of each simulation (Figure 10). For ore thickness, zero values were retained to reflect the estimated ore tonnage. However, for nodes having simulated thickness of zero, nickel grade values were set to null prior to reblocking so that the reblocked average would reflect the grade of the panel ore tonnage. Note that this is required because the input data only has an associated grade when the thickness is greater than zero.

The resulting distributions of the panel means were then interrogated to derive key statistics of the distributions. Due to the boundary constraint imposed on the study area, the peripheral 100 m square panels contained less than 100 simulation nodes. For these edge panels, the number of contained nodes captured was used to compute the panel proportion. Resource tonnages for each panel were calculated as the product of panel area, proportion, average reblocked thickness, and a density value of 1.5 t/m^3.

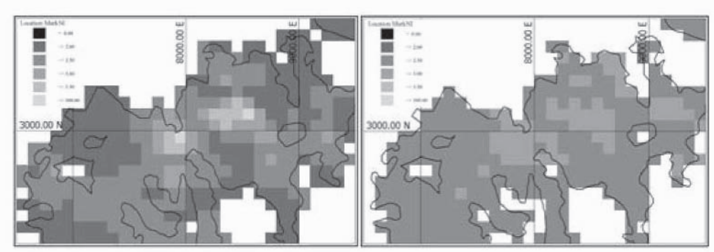


Figure 10 Panel averages for ore-thickness (left) and nickel grade (right)

The shape of the reblocked mean distributions for each panel is near normal, which is a feature consistent with the Central Limit Theorem of statistics. This theorem states that a distribution of means tends towards a normal distribution, as the number of samples used to compute the mean becomes large.

Confidence limits for the mean can be calculated using normal distribution theory. For example, the mean ± 1.645 standard deviations, contains 90% of the area under the standard normal distribution curve. The 90% confidence limits can be expressed relative to the mean of each panel distribution to give the relative 90% confidence limits (1.645 x panel standard deviation / panel mean). Specifically, for the reblock means of ore-thickness or nickel grade, the relative 90% confidence limits quantify the variability that can be expected (9 times out of 10) during production. For this study, the target acceptance threshold for Measured and Indicated resource classification was set to 90% confidence limits within ± 15% of the mean for a given production period, as discussed further below.

The relative 90% confidence limits were computed for each panel within the study area as shown in Figure 11. These plots confirm the intuitive conclusion that the lowest risk for both attributes occurs where the data spacing is closest (see Figure 2 for data locations). Additionally, the risk for ore-thickness (tonnage) is significantly higher than the risk for nickel grade, with the nickel grade meeting the benchmark of 90% confidence limits within ±15% of the mean for most of the Centre area on a panel-by-panel basis.

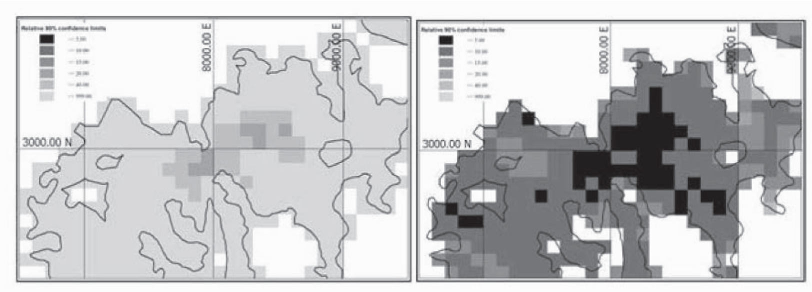

Relative 90% confidence limits
8000.00 E
9000.00 E
3000.00 N
Relative 90% confidence limits
8000.00 E
9000.00 E
3000.00 N

Figure 11 Panel relative 90% confidence limits for ore-thickness (left) and nickel grade (right)

One of the objectives of the simulation approach is to take into account the local variability of the data in assessing risk. Figure 12 shows two panels, A and B, with near identical data configurations in the centre of the study area, along with the ore thickness input data and the relative 90% confidence limits of panels within this area. The histograms of the simulation panel means for panels A and B are shown to the right of the data map. A kriging estimation variance approach to risk assessment would have given the same kriging variance and confidence limits for both panels. However, the relative 90% confidence limits using simulation are ± 30% for panel A and ± 20% for panel B.

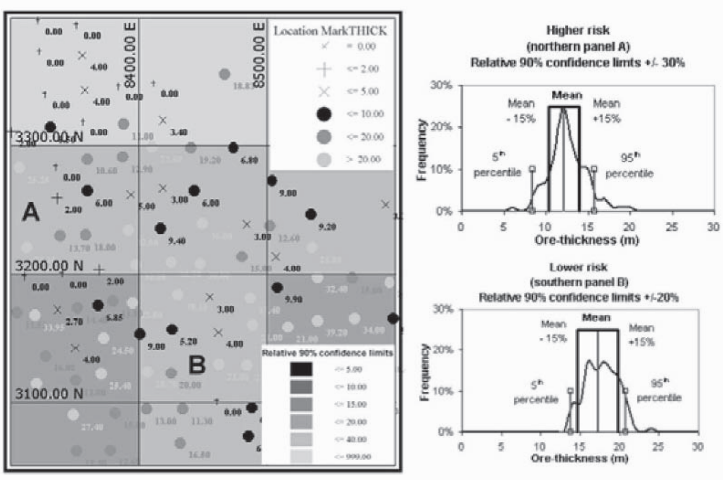
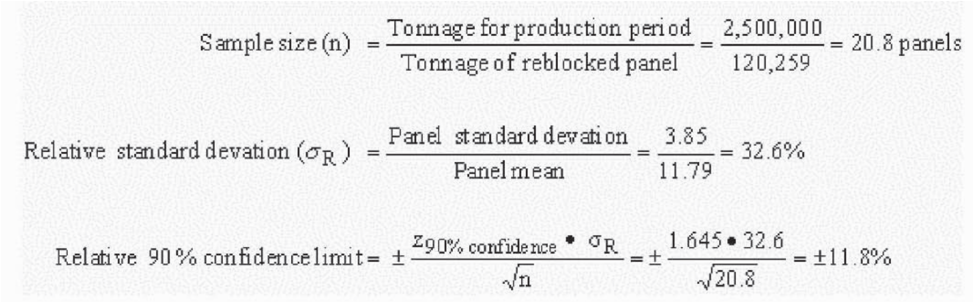


Figure 12 High (A) and low (B) risk panels and distributions of panels means for each panels; 5th and 95th percentiles are compared to interval ± 15% of the panel mean.

7 Production scaling of confidence limits

The relative errors computed on a panel-by-panel basis do not accommodate the fact that multiple panels will be mined during any mine production period. Because a production schedule is not yet available to allow reblocking or aggregation of panels to reflect actual production periods, it was assumed that a number of panels, n, of similar character (grade and/or depth and/or thickness) would be mined in a given production period. Further, the panels were assumed to be large enough to assume independence between panels. This assumption permits adoption of an approximation of the standard error formula ($\sigma/\sqrt{n}$) to compute the standard deviation for panels that are aggregated to represent a larger volume of mine production. For samples of size n from a large population, relative 90% confidence limits can be estimated by the product of the standard error of the mean and the standard normal deviate or z-value (1.645) of the confidence limit of interest.

For the purposes of this study, the size of the sample n is derived from the ratio of the production period tonnage to the tonnage within each reblocked panel. This formula assumes independence of realisations for each of the panels constituting a production period. The semivariograms show first ranges of less than a panel width (100 m) and secondary ranges of 200 m or less. Given these features and the fact that the majority of the study area is defined by 80 m spaced sampling or less, the assumption of independence was taken to be reasonable. [1] The example below shows this calculation for one panel in the study area using the assumption of a production rate of 2.5 Mt per annum and shows the relative 90% confidence limits are ± 11.8% per annum when multiple panels of the same risk character are scheduled to the meet the production requirement.



$$\text{Sample size}\,(n) = \frac{\text{Tonnage for production period}}{\text{Tonnage of reblocked panel}} = \frac{2{,}500{,}000}{120{,}259} = 20.8\ \text{panels}$$

$$\text{Relative standard devation}\,(\sigma_R) = \frac{\text{Panel standard devation}}{\text{Panel mean}} = \frac{3.85}{11.79} = 32.6\%$$

$$\text{Relative 90\% confidence limit} = \pm\frac{z_{90\%\ \text{confidence}} \bullet \sigma_R}{\sqrt{n}} = \pm\frac{1.645 \bullet 32.6}{\sqrt{20.8}} = \pm 11.8\%$$

Using this method the relative 90% confidence limits were computed for annual (2.5 Mt) and quarterly (0.625 Mt) production periods for ore-thickness.

The quarterly production (0.625 Mt) map of ore-thickness shows that most of the deposit area has relative 90% confidence limits exceeding ± 15% of the panel mean. Only areas of dense drilling and panels with low ore tonnage (generally thin or partially filled panels) are below this threshold. The thin ore areas have smaller confidence intervals, as a large number of "equivalent" panels are mined in the

[1] A data check of the semivariogram of residuals: [simulated reblocked values – the mean reblocked value for panels], showed slight dependencies at 100 m, and were the study to be repeated, a larger panel size would be chosen to reduce/remove these dependencies.

production period. This result is an undesirable artefact of the method; however the tonnage involved is small. On an annual basis (2.5 Mt), most of the study area has relative 90% confidence limits below the ± 15%ore-thickness target value. However, on a quarterly basis most of the study area has relative 90% confidence limits above the ± 15% ore-thickness target value.

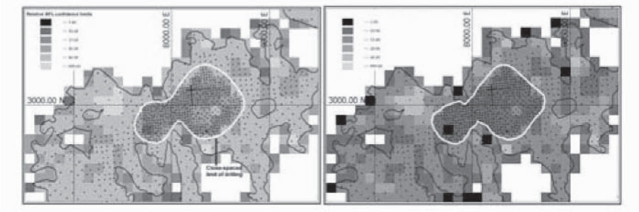


Figure 13 Ore-thickness relative 90% confidence limits, scaled to production rates of 0.625 Mt (left) and 2.5 Mt (right); point markers are drillhole collar locations

8 Discussion

From this study, Falconbridge concluded that the variability of nickel grade presented a low-risk. Wide spaced drilling (80 m spacing) adequately defined the nickel grade with relative 90% confidence limits within ± 15% of the mean for 100 m square panels. However, tonnage risk was considered to be high on a panel-by-panel basis. Rescaling the risk to quarterly and annual production periods revealed that the annual risk was acceptable, but that close-spaced drilling would be required to increase the confidence in tonnage to the target of relative 90% confidence limits within ±15% of the mean on a quarterly basis.

The 20% relative accuracy confidence limits are shown in Figure 13 (left panel) and enclose an area of dense drilling. The required 90% confidence limits for Measured Resources were eased to ± 20%, provided steps were taken to ameliorate the additional risk. The risk is ameliorated by implementing detailed drilling on a 10 m spacing a year in advance of production and increasing the number of mining faces (exposed ore) available for production.

This was a pioneering study for the authors and organisations involved. It was successful in that the initial study was completed rapidly, focused infill drilling requirements to reduce risk and assisted Falconbridge in managing the risk profile for a billion-dollar project.

9 References

Arik, A. (1999). An Alternative Approach to Resource Classification. 1999 APCOM Proceedings, SME. Denver. p 45-53.

Deutsch, C.V. and J ournel (1998). GSLIB GeostatisticalSoftware and User's Guide. Second Edition. Oxford Press. New York.

Yamamoto, Y.K (2001). Computation of global estimation variance in mineral deposits. In: Computer Application in the Minerals Industries, Xie, Wang & J iang (eds). Swets & Zietlinger. Lisse. p 61-65.

MINERAL RESOURCE CLASSIFICATION THROUGH CONDITIONAL SIMULATION

TOMASZ M. WAWRUCH and JORGE F. BETZHOLD
Vice-Presidency Mineral Resources,
Anglo American Chile, Santiago

Abstract. Through the Mineral Resource Classification the quantification of uncertainty/confidence on modelled geometry/estimates is to be addressed and established. Estimates represent different levels of reliability. The classification role is to ensure that the characteristics, quantity and quality of mineralised material is adequate for the proposed project or mining program, assuring the use of full plant capacity and optimisation of the mining and metallurgical performance down stream to the final products.

1 Introduction

The purpose of the paper is to discuss a method of Mineral Resource classification based on conditional simulation applicable at production scale.

Evaluation, classification and reconciliation are integral parts of the Mineral Resource management process. Drilling and sample collection together with QA/QC validation periodically update this process. The aforementioned system is "factor-dependent". Any sophisticated method applied to the modelling and evaluation is worthless if sample collection, preparation and chemical assays amongst other "factors" are not properly controlled.

Evaluation on its own tries to predict short, medium and long term scheduled mining output. It is worthwhile to look for the right methods to provide good tonnage and grade predictions to benefit the mineral resource management and mining program. As a result of the imprecision of evaluation, Mineral Resource modelling undergoes different magnitudes of discrepancies in comparing to the actual results. This is one reason for which Mineral Resource classification is required. It gives to competent persons, who are aware of the risks and consequences involved in inadequate mineral resource prediction, the opportunity to express a confidence concerning the estimates assigned to rocks to be mined. Monitoring the production results through the reconciliation process gives an important feedback in order to measure discrepancies, quantify errors and tune back the system if performance is unsatisfactory.

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 479-489.

Some characteristics the mineral resource classification should fulfil are as follows:

- Classification must be transparent and objective to what is to be achieved through the set of parameters defined integrally and accordingly to the geometry/grade modelling and interpolation strategy.
- The classification method must be reproducible and auditable based on quantifiable principles, applicable at the production scale.
- Mineral Resource classification should consider the production scale over a given period of time to be reconciled. This refers to the support size/estimation error relationship.
- Acceptability concerning the discrepancy between prediction and actual results should be explicitly established according to the nature of mineralization and the operational requirements.
- Through the Mineral Resource classification output, it should be possible to define areas where the confidence concerning geometry/estimates requires improvement. The increased confidence should be quantified as a function of money spent on drilling and sampling collection campaigns.

2 Classification methodology

The proposed Mineral Resource classification approach quantifies the confidence in the evaluation result. The reliability of the estimated tonnage/grade at location "x" is established and measured as a function of variance calculated through a given number of conditional simulations.

This Mineral Resource classification methodology is based on:

1. Multiple realisations of the sequential Gaussian conditional simulation. Other simulation routines and multi-realisation engines can also be considered (Deutsch, 2002; Goovaerts, 1997).
2. Calculation of the coefficient of variability for local mining unit
3. Change of support related to 1 to 3-monthly and yearly production panels
4. Estimation error according to established requirements in terms of % of error and % of confidence limit.

As an example for the Base Metals industry, widely accepted rules of Mineral Resources classification are as follows:

- Mineral resources are classified as ***Measured*** when the local estimate, whose variability is corrected to monthly to quarterly production units, is estimated within 15% error at a 90% confidence limit.

- Mineral resources are classified as ***Indicated*** when the local estimate, whose variability is corrected to yearly production units, is estimated within 15% error at a 90% confidence limit.

- The mineral resources that do not fulfil the aforementioned criteria are classified as ***Inferred.***

2.1 MULTIPLE REALISATION OF SEQUENTIAL GAUSSIAN SIMULATION

The objective of conditional simulation is to generate an equally probable set of realizations that account for proportion/distribution and spatial geometry/grade variability inferred from conditioning data at global/local scale. During the simulation, simulated nodes are visited in a random fashion. The conditioning is extended to all of the data available within a neighbourhood of the location being simulated, including the original data and all previously simulated values. Given that the estimated model is inferred from sample statistics that are uncertain because of limited number of samples, *the purpose is to provide the measure of uncertainty given by the differences between **N** alternative simulated values at location **x***. Different simulations impart different global statistics and spatial features on each realisation. In this way, it is possible to establish the spectrum of possible values at any location. (Deutsch, Journel-1998)

The number of realisations needed depends on how many are judged sufficient to model the uncertainty being addressed. A cumulative coefficient of variability (COV) from a low number of simulations has a variable behaviour, which stabilises as the number of simulations from which it is calculated, increases. After certain number of simulations, the oscillation of COV between succeeding simulated realisations stabilises. Taking this fact into account the number of simulations to define the uncertainty model is established.

The reliability of the simulations is checked by comparative analysis done in three ways:

- Reproduction of global statistical distribution
- Spatial grade variability/continuity model reproduction
- Local grade reproduction between conditioning values and sets of “N” simulations

To ensure that simulated realizations correctly reproduced the spatial theoretical grade variability/continuity model, simulated realizations are submitted to the spatial variability analysis of each given realization. Every single simulated realization is then checked using the same spatial variability formula and the same parameters as those used to get the variogram/correlogram experimental points of the raw variable.

The spectrum of spatial geometry/grade variability models along selected directions covers the interval of possible solutions. Considering the uncertainty of the conditioning database and spatial grade variability model, the family of “N” simulations is accepted as fairly representing the imposed spatial variability/continuity model.

2.2 CALCULATION OF COEFFICIENT OF VARIABILITY FOR LOCAL BLOCK

The coefficient of variability - COV, a dimensionless measure, defines the magnitude of deviation relative to the average. The whole mineralised domain is analysed. The measure of relative dispersion for each local block, based on “N” simulated values is calculated. The distribution of COV varies over the orebody as a function of the amount of conditioning data, local grade variability and the spatial grade variability/continuity model. In a densely sampled area, the expected variability among the multiple simulations of the variable is expected to be lower than in a poorly sampled part of the orebody. Nonetheless, even over the densely sampled areas, where the local grade variability is high, high values of the coefficient of variability can be expected.

2.3 CHANGE OF SUPPORT TO PANEL PRODUCTION

A meaningful way to classify Mineral Resources is to take account of the variability of local blocks within a bigger production volume. The change from local volume/grade variability to the variability of the production panel tends to ensure that the expected metal contained within monthly and annual production units is estimated with an error not greater than the established tolerance at given confidence limit.

The way to express local geometry/grade variability as a variability of bigger mining units is through the variability reduction factor (Isaaks, Srivastava-1989). The calculation takes into account averages of COV computed for local mining blocks and production panels.

A series of author’s exercises with different kinds of deposits have shown that global monthly or annual variability reduction factors are not the best values to apply for Mineral Resource Classification purposes. Spatial COV maps usually show different local configurations for different geological situations. Using one global correction value increases the uncertainty for low variability areas and gives more confidence to high variability areas. It is considered that a more objective way to do this is by introducing local variance reduction factors.

It is proposed to divide the block population into classes of COV. The number of classes depends on what even-frequency per class is targeted. The classification is based on calculations computed within each group. Assuming that “n” local COV classes are analysed, “n” variability reduction factors are computed.

A randomly positioned production scale panel allows to accumulate a number of observations considered sufficient to calculate the COV for a given production scale. Following this, the variance reduction factor for monthly and annual production units is computed and applied to the local COV. In this way the local variability is corrected to production scale variability as if it was a fraction of bigger support.

$$f_R = COV_{\text{panel production}} / COV_{\text{Local}}$$

f_R - variability reduction factor
COV_{Local} - local coefficient of variability

Confidence limits are addressed by the mean of formula where the local coefficient of variability COV is corrected by variance reduction factor.

$$f_R * COV_{Local} * Z_C \leq 15\%$$

Where,

Z_C -confidence limit coefficient; accepted Normal distribution (0,1) if local COV is reasonably Gaussian

2.4 CLASSIFICATION CURVES

Having obtained the following:

- Percentile proportion of each class
- Local average of COV for each class
- Two COV for monthly and annual production panels corrected by variability
- reduction factor,

it is possible to visualise three curves called, for Mineral Resource Classification purposes, **Classification Curves**.

Pairs obtained from the class central point and average local COV are used to construct the local COV curve that increases as a function of increasing percentile population.

The measured and indicated classification curves are based on calculations computed within each group. The common feature of them is their decreasing nature, throughout the increasing percentile population.

The three curves create two intersections if Measured, Indicated and Inferred Mineral Resources are present. The curve at lower local relative variability represents the partition between Measured and Indicated Mineral Resources. The second decreasing curve at higher local variability establishes the limit between Indicated and Inferred Mineral Resources. At each intersection the proportion of Mineral Resources belonging to one of three Mineral Resources confidence classes can be read. Their Y-axis equivalents establish the separation thresholds between Measured, Indicated and Inferred Mineral Resources in terms of local COV.

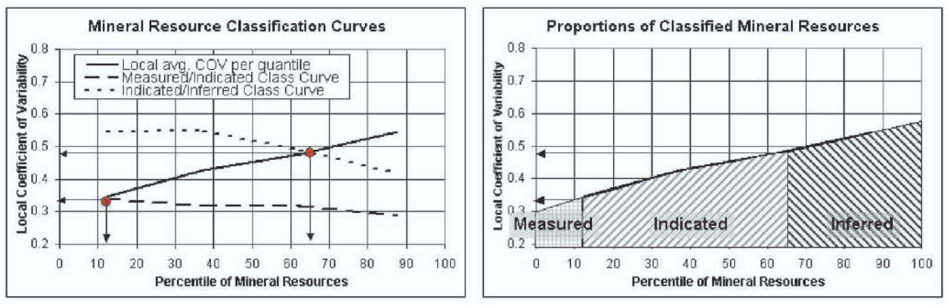


Figure 1. Graphic representation of Mineral Resource classification results

Based on the local COV at production support (monthly and annual) it is possible that one or both intersection points are not present. If the spatial continuity, sampling density and local variability are unfavourable, the proportion of Measured or Indicated Mineral Resources will not be present or will show a low proportion of the total.

3 Sensitivity of the proposed methodology

3.1 HISTOGRAM GRADE REPRODUCTION

Global/local grade distribution – since the coefficient of variability is used as a measure of local grade dispersion, any oversight of global or local grade distribution distorts the local relative variability distribution. As a consequence might be, underestimated local grades at assumed correctly reproduced spatial variability/continuity model increase values of coefficient of variability. This increases the amount of Mineral Resources at lower confidence class, which correctly classified would have been assigned a better category.

3.2 SPATIAL GRADE VARIABILITY REPRODUCTION

Spatial grade variability/continuity model – if incorrectly reproduced and accepted, it could significantly change the classified Mineral Resource proportions. If the spatial variability range of simulated realizations is too long, it produces a false effect of better continuity. The improved spatial grade continuity implies relatively higher dispersion variance distribution for the given production reference unit. As a result, more Mineral Resources can be classified with higher tonnage/grade uncertainty. The effect is predominantly pronounced over areas with scarce conditioning data (Figure 2, Table 1).

Short continuity ranges promote a faster decrease in variability whilst changing support from local increments to panel production units. Small values of the variability reduction factor increase the proportions of Indicated Mineral Resources.

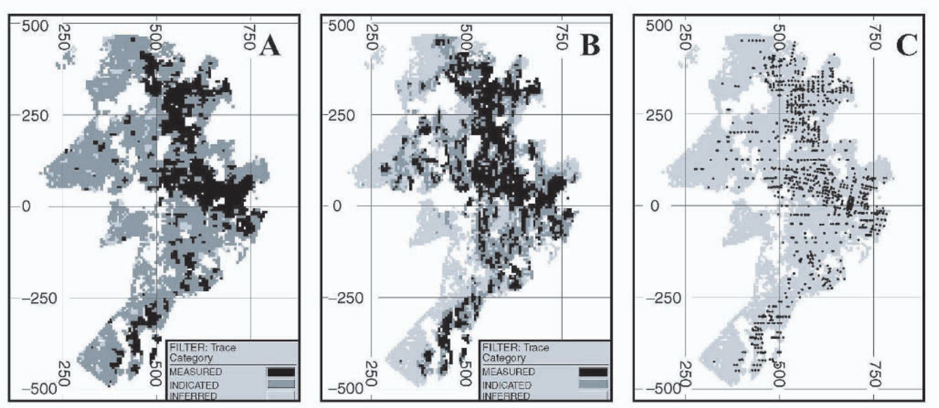


Figure 2. Distributions of Classified Mineral Resources as a function of spatial variability range reproduction: A: short-range, B: long-range, C: spatial data location

	Classified Mineral Resources – Proportions [%]		
Case	Measured	Indicated	Inferred
A: Short range	29.4	65.8	4.8
B: Long-range	29.7	28.6	41.7

Table 1. Proportions of Classified Mineral Resources as a function of spatial variability range reproduction. Stable % proportion of Measured Resources shows the data-driven effect and variable % proportion between Indicated/Inferred Resources reveals the model-driven results.

3.3 SIZE OF PRODUCTION PANEL

The concept of production panels may have a real or theoretical aspect. In the case of having a mining program for a given period of time, programmed areas/volumes can be used to calculate the COV of panel production as increments of local COV. This is a case to verify and quantify the uncertainty on tonnage/grade for the existing extraction program.

Without a mining program (project, pre-feasibility study) a theoretical approach to calculate variability for a production period can be used. Vertical dimensions can be taken from a possible number of benches envisaged in the production program. This can be estimated by comparing to other mines/projects of a similar nature. To scale up the horizontal dimensions, the character of spatial variability/continuity model (isotropy, anisotropy) and distribution of existing or assumed opened mineralised faces/stopes are decisive factors. Once established, the size of production panel can be an object for sensitivity study.

3.4 NUMBER OF LOCAL BLOCKS WITHIN PRODUCTION PANEL

Different orebodies present their proper geometric particularity. Whilst computing the statistics on local COV within monthly or annual production panels, some of the panels over the borders of the domain gather only a small number of local COV´s. To ensure the robustness of statistics, panels with small numbers of local blocks should be discarded.

3.5 FREQUENCY OF LOCAL BLOCKS PER CLASS

The suggested number of classes for the local COV should be between 4 and 10. This means that a small orebody could not be divided into an elevated number of classes having too few local blocks per class.

Sensitivity analysis on the aforementioned issues should be carried out. The output for the exercise is a family of classification curves for Measured/Indicated and Indicated/Inferred Mineral Resources. Through them the uncertainty of the Category proportions for the Mineral Resource is assessed. The average of computed answers is accepted to break-up the classified Mineral Resource proportions.

4 Quantified confidence improvement - example

This section shows the application of the discussed methodology. Together with the classification method, the reconciliation between predicted and actual proportions of classified Mineral Resources is presented.

The classification method applied Sequential Gaussian Conditional Simulation as engine to create conditioned, equally probable grade distributions. The local variability was expressed through the change of support for monthly and annual production panels. The criterion of an error within 15% at 90% confidence limit was used.

The orebody had been intercepted by 98 boreholes. The average distance between them was greater than 50m. Ranges of the spatial grade variability model were less than 50m and represented only subtle geometrical anisotropy. Following the procedure discussed in this paper, an initial classification of Mineral Resources has shown no Mineral Resources classified as Measured and only 12% of Indicated Mineral Resources (Figure 2). The quantity of Measured/Indicated resources indicated a need for new information to improve the confidence in geological resources.

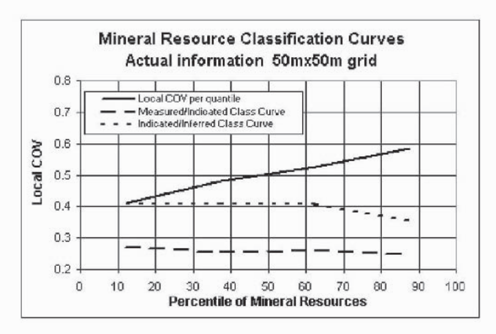


Figure 2. Graphic assessment of Mineral Resource classification

An exercise was carried out to quantify "a priori" a possible amount of Mineral Resources to be upgraded as a function of new drilling information. This was done by simulating virtual drilling campaigns. To optimise the drilling program, different sets of drilling grids were analysed.

It was assumed that in spite of only 346 samples from 98 boreholes regularly distributed over the orebody, the average grade and variance would not change drastically as a result of new data collection. This assumption has been assessed using the set of 51 simulations generated for the purpose of classification. At a 90% confidence limit the expected discrepancy concerning average grade was defined as 5.4% and variance as 6.8%.

Following this, the discussed classification methodology was applied. It was concluded that among multiple exploration strategies, a sampling grid of approximately 30m x 20m would allow to have 13% of Measured and 52% of Indicated Mineral Resources (Figure 3).

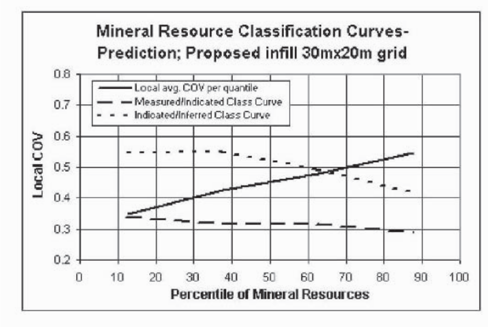


Figure 3. Prediction of Mineral Resource classification proportions based on virtual exercise

The actual exploration drilling program contains 72 new boreholes. In total 729 samples were conditioning the geometry and the grade distribution within the orebody. The uncertainty on geometry was assessed through probabilistic models and grade estimates within the orebody were reproduced using sequential conditional Gaussian simulation.

The classification methodology applied to Mineral Resources assigned the confidence level in the following (Figure 4) proportions: Measured Resources 8%, Indicated Resources 60% and Inferred Resources 32%.

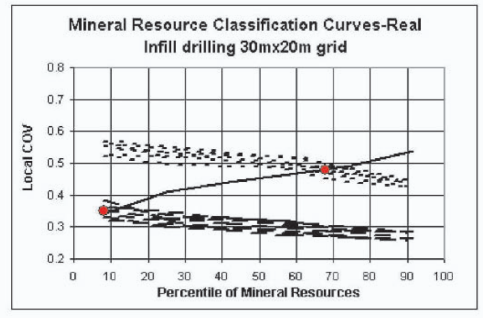


Figure 4. Mineral Resource classification based on the updated database

The targeted proportion of classified Mineral Resources had been reached. The use of a family of classification curves allowed to assess the uncertainty concerning the confidence on Mineral Resource classified proportions.

5 Conclusions

Mineral Resource classification is an integral part of Mineral Resource evaluation and reconciliation. It constitutes an important strategic tool allowing to assess tonnage/grade uncertainty for the mining program.

A golden formula to classify mineral deposits does not exist. Different methods to express confidence in Mineral Resource evaluation are employed. Although the common practices have been developed, the robust approach toward the Mineral Resource classification method through the uncertainty quantification is not always exercised.

The method presented in this paper proposes to quantify confidence through equally probable, spatially conditioned multi-realizations. As an engine to create the conditional spatial grade distribution the sequential conditional Gaussian simulation was used. The classification approach integrates transparency, objectivity and geostatistical tools commonly used in modelling and evaluation. The relevant issue is to express the uncertainty of Mineral Resources as a function of production panels that is to be reconciled for a determined production period. The reproducibility of the classification method is achieved through parameters defined numerically. This makes the classification method easily auditable. Classification curves allow visualising the classification output.
This classification approach includes quantification of confidence in estimated Mineral Resources. It can be applied to geological projects and mining operations. The results

are submitted to a continuous monitoring and validation process through reconciliation figures on a monthly basis.

Acknowledgements

We would like to acknowledge Anglo American Chile for supporting this research.

References

Deutsch, C.V., *Geostatistical Reservoir Modelling*, Oxford University Press, 2002.

Deutsch, C.V. and Journel, A.G., *GSLIB: Geostatistical Software Library and User's Guide*, 2nd Edition, Oxford University Press, 1998.

Goovaerts, P., *Geostatistics for Natural Resources Evaluation*, Oxford University Press, 1997.

Isaaks, E.H. and Srivastava, M.R., *An Introduction to Applied Geostatistics*, Oxford University Press, 1989.

This page intentionally blank

GEOSTATISTICAL INVESTIGATION OF ELEMENTAL ENRICHMENT IN HYDROTHERMAL MINERAL DEPOSITS

ABANI R. SAMAL[1]* AND RICHARD H. FIFAREK[1] AND RAJA R. SENGUPTA[2]
[1]Dept. of Geology, Southern Illinois University, Carbondale, IL 62901-4324, USA
[2] Dept. of Geography, McGill University, 805 Sherbrooke St. W.
Montreal, Quebec, Canada H3A 2K6
**Email: arsamal@yahoo.com*

Abstract. Hydrothermal hypogene processes enrich or deplete rocks in specific suites of elements to form mineral deposits. Subsequent geochemical processes, such as near surface oxidation, commonly remobilize previously developed element associations. Late stage oxidizing fluids and elemental enrichment/depletion are commonly guided by permeable geologic structures in the deposit. An investigation of element redistribution is possible using a cross-covariance analysis between pairs of elements. The maximum positive cross-covariance of a pair of variables yields a vector, known as the lag vector. This lag vector may indicate the direction and distance of element displacement from their original loci.

This paper discusses the application of cross-covariance in modeling the anisotropy of metal redistribution as a function of late oxidation in the Pierina hydrothermal Au-Ag deposit. Cross-covariance analyses of assay values from drill-hole samples in both the oxidized and unoxidized zones are calculated and lag vectors [$l_{xy(O),}$ where x and y are elements in a zone O] are derived to infer a preferred path of metal remobilization. The azimuths of lag vectors for the element pairs Ag-Au, Cu-Au and Cu-Ag in the oxidized zone correspond to the orientations of recognized faults and fractures in the deposit. This implies that the remobilization of Au, Ag and Cu by oxidizing fluids was strongly controlled by specific fault or joint sets. The data for all three element pairs from the unoxidized zone suggested structural controls different from those of the oxidized zone.

These results imply that a cross-covariance analysis for pairs of elements may be used to infer structural controls on fluid flow which might be responsible for element remobilization and possible enrichment. Such an analysis may be useful in mineral exploration for predicting metal enrichment and the location of exotic (transported) deposits.

1. INTRODUCTION AND BACKGROUND INFORMATION:

The spatial pattern of element distribution in hydrothermal deposits results from the overprinting of multiple alteration events. The late stage, near-surface oxidation of deposits is related to vertical movements of the groundwater table, and typically results in the downward transportation of elements, possibly with some lateral component of

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 491-500.

movement. Fluid flow and element mobilization are guided by major structural features of high permeability (faults, joints, etc.). Oxidation is important to the economics of mining large, low-grade, disseminated metal deposits in that it releases metals encapsulated in sulfides, thereby making the ore amenable to low-cost extraction technologies, and in the enrichment of metal grades.

Element distributions in mineral deposits can be modeled numerically through the application of geostatistics. A multivariate geostatistical analysis using maximum cross-covariance values can be applied to model the spatial dependency of two metals and to relate the results to orientations of mineralized faults and joints (Samal and Fifarek, 2003). In this study, we explore the application of cross-covariance analysis to assay data from the Pierina (Peru) Au-Ag deposit where late stage oxidation has clearly remobilized metals.

1.1 Cross-covariance and lag effect:

With the assumption of second-order stationarity and ergodicity, the covariance ($C_{i(h)}$) of any variable *i* measures spatial dependency of the *same* variable (or values of the same property of material) at two locations, where *h* is the distance of separation (a vector) between the two locations. Similarly, under assumption of joint second-order stationarity, the cross-covariance function $C_{ij(h)}$ measures the spatial dependency between *two* variables *i* and *j*, here concentrations of two elements, separated by vector *h*. The cross-covariance between two elemental concentrations *i and j* is expressed as (Equation 1):

$$C_{ij\,(h)} = \frac{1}{n}\sum (i_h - m_i)(j_{-h} - m_j) \tag{1}$$

where m_i is the mean of the variable *i* and m_j is the mean of the variable *j*

The cross-covariance analysis is not an even function (Wackernagel 1998). The asymmetric behavior of the cross-covariance function between two variables in isotopic, heterotopic or partially heterotopic datasets is seen in the assay values of gold deposits. The dataset used in our cross-covariance analysis is partially heterotopic, i.e. data for all variables are not available for all sample locations. The asymmetry can be defined as $C_{ij(h)} \neq C_{ij(-h)}$, where $C_{ij(-h)}$ is the cross-covariance of *i* and *j* separated by a distance *h* but in the opposite direction. But if both the sequence of variables and the sign of the lag (*h*) are changed, the value of the cross-covariance function $C_{ij(h)} = C_{ji(-h)}$ (Wackernagel, 1998; Isaaks and Srivastava, 1989). Cross-covariance values can be positive when the variables at the end points of the *h* vector are on the same side of their means, i.e, $i_h > m_i$ and $j_{-h} > m_j$ or, $i_h < m_j$ and $j_{-h} < m_j$; where i_h is the value of *i* at the head of the vector *h*, and j_{-h} is the value of *j* at the tail of the vector *h*. Depending on how far they are from their respective means, the value of a positive cross-covariance will be high or low. So, if *i* and *j* are extremely high values (enrichment of both elements) or extremely low values (depletion of both elements), both cross-covariance values will be positive and high (not low). But if one element is enriched and the other element is depleted, then the cross-covariance is negative.

The lag effect is the vector ($\boldsymbol{l_{ij}}$) separating the locations of extreme values of two variables, which in some geological environments may be due to the delay in enrichment of one element with respect to the other at two different locations (Isaaks and Srivastava, 1989, Goovaerts, 1997). This offset distance is also termed the delay effect when time-series data is considered, such as in most environmental applications (Wackernagel, 1998). In the oxidized zone of a hydrothermal mineral deposit, the offset between the concentrations of two elements is due to differences in their mobility resulting in the enrichment and depletion of different elements at different locations. In a preliminary exploratory study of the Pierina deposit (Samal and Fifarek, 2003, Samal, Fifarek, Sengupta 2003 and Samal, Fifarek, Mohanty 2004), lag vectors were derived that generally corresponded to the orientation of specific major fault and joint systems. Deriving the lag-vectors from maximum positive cross-covariance values (maximum values in any direction) ignores other higher cross-covariance values. In this paper other high cross-covariance values are taken into consideration for three pairs of elements (Ag-Au, Cu-Au, and Cu-Ag) in the Pierina deposit.

1.2 Deposit geology:

The Pierina deposit is located in the Ancash Province of Peru. It is a world class, high sulfidation, epithermal Au-Ag deposit with anomalous but uneconomic concentrations of Cu, Zn, As and Hg. The geology and genesis of the deposit are presented by Fifarek and Rye (in press), from which the following summary is taken.

The Au-Ag ore-body is sub horizontal, elongates N-S, and almost entirely hosted by rhyolite ash flow tuffs that overlie porphyritic andesite and dacite lavas and are adjacent to a crosscutting and interfingering dacite flow dome complex. Alteration and mineralization occurred 14.5 Ma ago as a result of the expulsion of fluids, gasses and metals from an underlying magma. Highly acidic fluids formed at the level of the deposit where slowly rising magmatic vapors condensed and mixed with cool meteoric groundwater. The progressive neutralization of these migrating acid-sulfate fluids led to zoned alteration assemblages from proximal vuggy quartz through quartz-alunite ± clay and intermediate argillic to distal propylitic. Copper-gold-silver mineralization largely followed alteration and is marked by the deposition of enargite (Cu_3AsS_4), electrum (Au-Ag), acanthite (Ag_2S) and related minerals. The primary elemental associations and concentrations in the sulfide deposit were established at this time.

A late oxidizing event related to a near-surface, steam-heated process was superimposed on the deposit during the waning stage of hydrothermal activity that was accompanied by a drop in the water table. These oxidizing fluids led to the destruction of sulfides and the formation of barite, hematite, goethite and minor jarosite. Consequently, the previously established elemental concentrations and associations were substantially modified due to the remobilization of most elements. Late oxidizing fluids pervaded rocks of the upper 200 to 300 m of the deposit and particularly followed open faults and joint sets.

Exploration drilling on mostly 50 m centers and assays of 1 m intervals of drill core or cuttings provided an extensive database of Cu, Au and Ag values. Additional datasets were generated from information on the distribution of alteration and fracture-filling minerals in the exploration drill holes. Together, the datasets constitute the basis for evaluating the remobilization of metals (Au, Ag and Cu) by oxidizing fluids. Approximately 24,500 data points were selected from the oxidized zone and approximately 14400 samples were selected from the unoxidized zone, by excluding widely spaced data points.

2. DATA ANALYSIS:

2.1 Data Preparation:

The drill-hole data were visually examined in a 3D environment using GEMCOM and Arc-GIS software-systems. For analytical purposes, a single table with records of Au, Ag and Cu and alteration details was created within GEMCOM. The oxidized zone is characterized by the presence of iron-oxides (FeOx) whereas the unoxidized part of the deposit is marked by the absence of FeOx. A solid model was then created for the alteration.

Using GEMCOM, two tables of data formatted for the geostatistical software ISATIS geostatistical software were created: one for the oxidized zone and the other for the unoxidized zone. A selected portion of the data was chosen from the area of regularly spaced drill-holes for analysis.

2.2 Geostatistical Analysis:

A univariate variography (covariance) analysis was used to model the anisotropy of individual elements in this mineral deposit. For pairs of elements, a cross-covariance analysis was used to derive lag vectors. The cross-variogram is an even function (Wackernagel, 1998, p 147) that fails to detect anisotropy and therefore is not relevant to this study.

ISATIS was utilized to analyze for the cross-covariance of the three variables, Au, Ag and Cu. Each set of data, oxidized or unoxidized, was analysed in 62 directions to cover all possible directions in 3D space with a 30° angular tolerance for each direction. Out of these 62 directions, 12 directions were on the horizontal plane (reference plane) and two in the vertical plane (up and down). The remaining 48 directions are defined in 3D space as 4 directions in 12 vertical planes whereby each plane includes one horizontal direction. On each plane, these 4 directions are separated by 30° between the horizontal and vertical directions. A 50m lag was chosen for all directions except the vertical directions where the lag distance was set as 10m.

The cross-covariance for Ag-Au, Cu-Au, and Cu-Ag pairs was calculated using the exploratory data analysis tool of ISATIS. It is noteworthy that ISATIS calculates the cross-covariance in each specified direction and its reverse direction, in other words, the ISATIS software calculates $C_{ij}(h)$ and $C_{ij}(-h)$. A cross-correlation analysis of the same pair of variables taken in the same sequence was performed in order to cross-check the results of the cross-covariance analysis. The cross-correlation (CC_{ij}) function (Equation 2) is:

$$CC_{\mathrm{ij\,(h)}} = \frac{1}{n}\sum\frac{(i_h - m_i)(j_{-h} - m_j)}{\sigma_i\sigma_j}. \tag{2}$$

Experimental cross-covariograms for each direction were plotted for a comparative analysis made in two ways: 1) for each pair of variables, high positive cross-covariance values were ranked and the corresponding directions were compared with the results from the previous study (Samal & Fifarek, 2003), and 2) the directions were compared with recognized structural trends. For reasons of clarity only those directions with the top three cross-covariogram values are shown in the following figures.

3. RESULTS AND DISCUSSIONS OXIDIZED ZONE:

Among the three variables, Cu in the oxidized zone has the highest variance followed by Ag and Au (Table 1). This reflects the relatively wide range of Cu values and suggests that the leaching of Cu was more extensive and the element more mobile than Au and Ag.

Table 1: General statistics of variables: oxidized zone

	Mean	Variance	Covariance	
Au	1.4 ppm	14.3 ppm^2	Ag & Au	61.4 ppm^2
Ag	12.7 ppm	1493.3 ppm^2	Cu & Au	128.7 ppm^2
Cu	128.7 ppm	139047.4 ppm^2	Cu & Ag	980.7 ppm^2

From a comparison of covariogram plots (not shown) of the three variables it is clear that the spatial dependency of Au and Ag is very similar. The covariance of Au and that of Ag at shorter distances of separation exhibit very high values along ENE to ESE directions (azimuths of 60° , 90° , 120°) at a lag-interval of less than 50m. The ranges of 150m (approximate) are higher along these directions than in other directions (e.g., azimuths of 0°). Additionally, the covariance values fall rapidly to low values along generally North-South directions.

The cross-covariograms of Ag and Au (Fig. 1) indicate the lag vector ($l_{AgAu(O)}$) is oriented along an East-West to ENE-SSE (Azimuth 90° & 60°) directions with a shallow dip of 30° (±15°) toward East. Experimental cross-correlogram plots of Ag-Au produce a similar pattern as that of the cross-covariograms (Fig. 2). The sequence of maximum to lower cross-correlogram values is along the same directions (Azimuth 90° & 60° and dip of 30° (±15°) toward East) as seen in the cross-covariograms.
The cross-covariograms of Ag and Cu yield a preferred lag vector ($l_{AgCu(O)}$) of azimuth 60°, dip 30°, followed by vectors with azimuth 270°, dip 60° and azimuth 240°, dip 60° (Figure 3). It can be inferred that, with respect to Cu enrichment (or depletion), a significant lateral movement of Ag has occurred in ENE-WSW to E-W directions. With the angular tolerance (30°) used in the analysis, it is likely that elemental remobilization is controlled by joints and faults aligned approximately ENE-WSW to E-W directions.

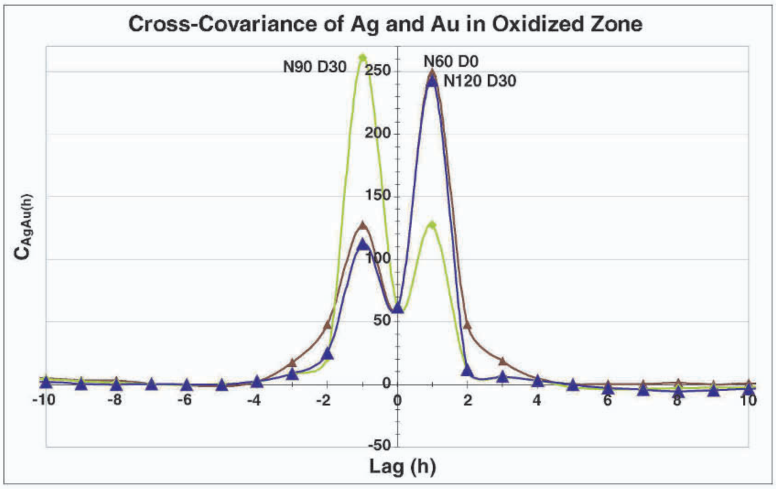


Figure 1. Cross-covariogram plots of Ag-Au in the oxidized zone (azimuths and dips shown for three highest values).

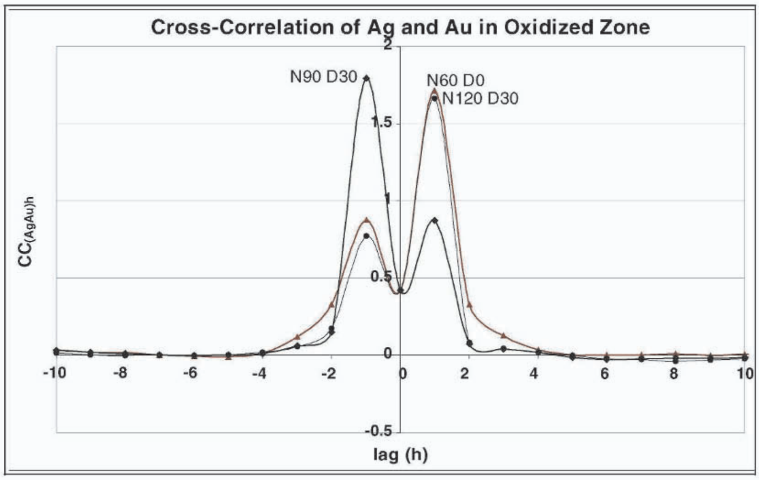


Figure 2. Cross-correlation plots of Ag-Au in the oxidized zone (azimuths and dips shown for three highest values).

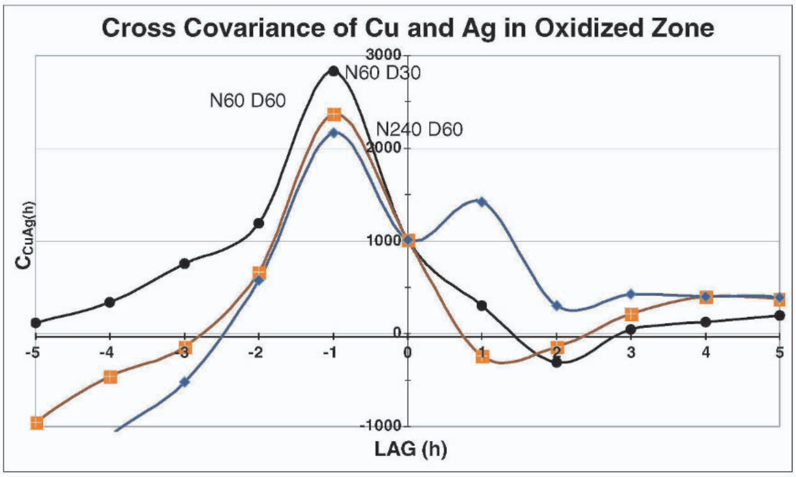


Figure 3. Cross-covariance of Cu – Ag in Oxidized zone (azimuths and dips shown for three highest values).

The cross-covariance plots of Cu and Au indicate a $l_{CuAu(O)}$ of azimuth 150°, dip 60°. Other prominent cross-covariance values imply vectors with azimuths of 270° dip 60° and 240°, dip 60° (Figure 4). From these observations, it is evident that, with respect to copper, the fluid transport and enrichment/depletion of gold and silver was in a general ENE-WSW to East-West direction.

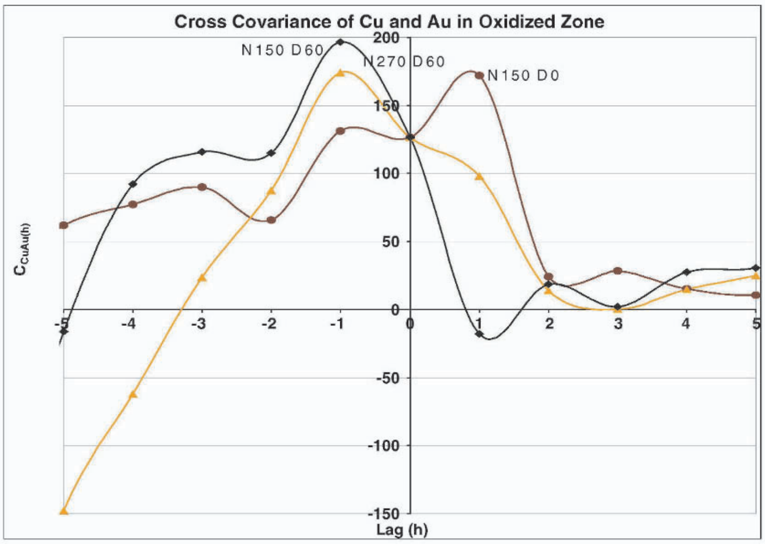


Figure 4. Cross-covariance of Cu and Au in Oxidized zone (azimuths and dips shown for three highest values).

In summation, Ag and Au have a similar spatial dependency pattern (as documented by their covariance values). The cross-covariance patterns of pairs of metallic elements (Ag-Au, Cu-Ag and Cu-Au) are useful in deriving vectors, which implies the shallow dipping transport of Au with respect to Ag along East, SSE and SSW directions. The cross-covariogram plots of Au-Cu and Ag-Cu pairs also suggest general West, East to ENE directions of fluid flow. Orientations of the major structural trends (joints and faults) that guided oxidizing fluids, as identified in this study, are summarized in Table 2. These are ENE to ESE and WSW to WNW directions, which are common in the cross-covariance analysis of all three pairs.

Table 2. Summary of prominent cross-covariance values

ANALYSIS	Rank	Representative directions	Comments
Cross-Covariance (Au & Ag)	1	Azimuth 90º and Dip 30º	Mostly ENE-ESE to WSW-WNW directions and shallow dips of 30° to 60° (±30º)
	2	Azimuth 240º	
	3	Azimuth 120º and Dip 30º(-*h*)	
Cross-Covariance (Ag & Cu)	1	Azimuth 60º and Dip 30º	
	2	Azimuth 60º and Dip 60º	
	3	Azimuth 240º and Dip 60º	
Cross-Covariance (Au & Cu)	1	Azimuth 150º and Dip 60º	
	2	Azimuth 270º and Dip 60º	
	3	Azimuth 150º	

3.1 Unoxidized Zone:

Data for the unoxidized zone were treated in the same manner as data for the oxidized zone. The unoxidized zone lies below the oxidized zone and is represented by fewer assays relative to the oxidized zone. For Ag - Au pairs, the lag vector ($l_{AgAu(U)}$) has an azimuth of 120° and dip of 30° (Fig 5). Other prominent directions of possible Au and

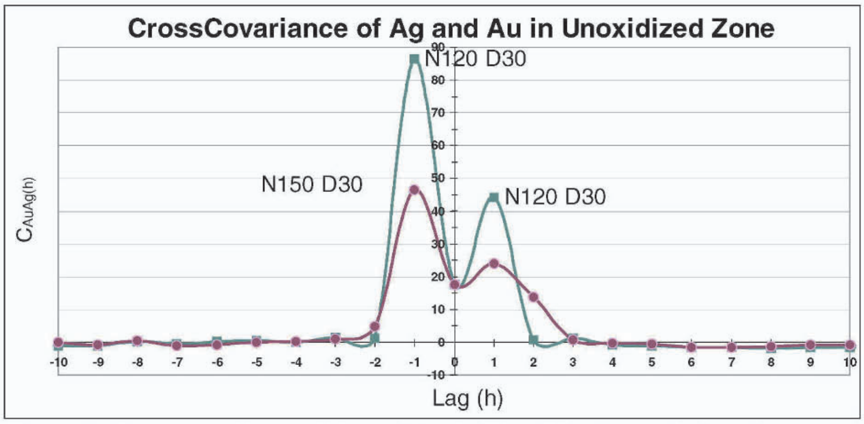


Figure 5. Cross-covariance of Ag and Au in Unoxidized zone (azimuths and dips shown for three highest values).

Ag movement during hydrothermal activity are along azimuth 300° , dip 30° and 150° , dip 30° . For Cu and Au pairs, the lag vector $l_{(AuCu(U))}$ has an azimuth of 90° , dip 60° , and a lag distance of separation of less than 50m followed by azimuths of 120° , dip 60° , and 270° , dip 60° (Fig 6). The analysis for Agnad Cu pairs suggests no preferred orientation of metal separation and fluid flow.

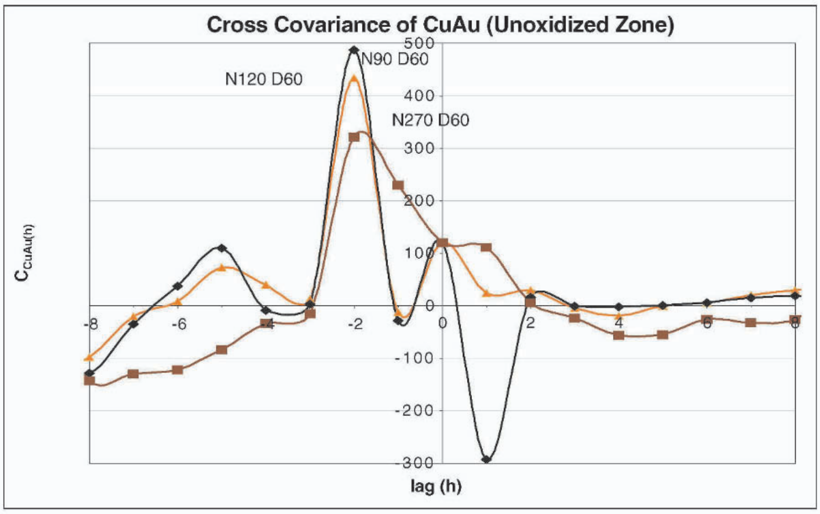


Figure 6. Cross-covariance of Cu and Au in Unoxidized zone (azimuths and dips shown for three highest values).

The azimuth 120° is common to Ag-Au and Au-Cu pairs and along which cross-covariance values are sufficiently high to suggest a prominent geologic trend. With an angular tolerance of 30° , the major directions of elemental remobilization are along ESE to east. This direction may represent a set of vectors along a set of fault or joint planes that are of pre-oxidation age. A structural study of mine exposures revealed a prominent set of faults and joints along this trend, as well as the other directions inferred (azimuths 120° , 90° , 240° , 270° & 300°) from this cross-covariance study.

4. CONCLUSIONS:

Based on the above observations, the following conclusions are possible.

i. A covariogram analysis suggests that Ag and Au have similar spatial patterns of distribution that differ from that of Cu in the oxidized zone of the Pierina deposit. Both elements show high covariance (a measure of spatial dependency) values at distances of separation less than 50m, whereas Cu shows no preferred lateral orientation of spatial dependency. Cross-covariance analysis of Ag & Au, and these two elements paired with Cu suggest downward and lateral mobilization of elements.

ii. In both oxidized and unoxidized zones, the general orientation of elemental mobilization is inferred to be along ENE, East, ESE, WSW and West directions with an angular tolerance of 30° . The major mineralized joints and faults in the Pierina gold deposit are oriented along these directions.

iii. Overall, the multivariate cross-covariogram analysis derived vectors of metal separation that coincide with recognized trends of major faults and joint sets in the Pierina deposit. Consequently, this type of analysis may be generally applied to hydrothermal mineral deposits as a means of identifying the structural features that guided hydrothermal and particularly oxidizing fluids resulting in the deposition and subsequent mobilization of metals.

Further research directions are suggested to refine and verify the validity of these results. These directions include the following:

- The odd parts of the cross-covariance (Goovaerts, 1997, p 73; Wackernagel, 1998, p 147; Webster and Oliver, 2000, p196) add to the anisotropic behavior of the results, whereas even parts of the cross-covariance are isotropic. It may be useful to model the odd parts of the cross-covariance and derive the lag-vectors for different pairs of the variables from the maximum and other high values.
- Using a geochemically identified immobile element in the pairs to better quantify distances of element separation and the location of enrichment zones.

5. ACKNOWLEDGEMENTS

The authors thankfully acknowledge Barrick Gold Company for providing the assay data used in this study and their support of geologic studies of the Pierina deposit. We thank Dr. Pierre Goovaerts for his valuable suggestions during the early stage of this research work.

6. REFERENCES:

Fifarek, R.H., and Rye, R.O., 2004, “Stable Isotope Geochemistry of the Pierina High Sulfidation Au- Ag Deposit, Peru: Influence of Hydrodynamics on SO_4^{2-}-H_2S Isotopic Exchange in Magmatic Steam and Steam-Heated Environments. Chemical Geology”

Goovaerts, P., “Geostatistics for natural resources evaluation”, Oxford University press, 1997

Isaaks, E.H. and Srivastava, R.M. “An Introduction to Applied Geostatistics”. Oxford University Press, New York. 561 p. 1989.

Samal. A. R, and Fifarek. R. H., 2003, “Application of Cross-Covariance in Geostatistical Modeling of Elemental Remobilization in Hydrothermal Mineral Deposits” in IAMG 2003 Conference Proceedings, Portsmouth, England, 6 pp.

Samal. A. R, and Fifarek. R. H., Sengupta. R, 2003, “Geostatistical modeling of elemental remobilization in hydrothermal deposits” in Abstratcts with programs, Vol 35, No 6, GSA annual meeting, Seattle, 2003

Samal. A. R., Fifarek. R. H. Mohanty, M. K., 2004, “Spatial modeling of elemental mobilizationin a hydrothermal gold deposit” in Abstratcts with programs, GSA North-Central Section, Annual meeting, St. Louis, 2004

Wackernagel Multivariate Geostatistics, Second Edition, Springer-Verlag Heidelberg, p. 145, 1998

Webster. R, Oliver. M., “Geostatistics for Environmental Sientists, J ohn Eiley & Sons, Ltd, 2000

VALUING A MINE AS A PORTFOLIO OF EUROPEAN CALL OPTIONS: THE EFFECT OF GEOLOGICAL UNCERTAINTY AND IMPLICATIONS FOR STRATEGIC PLANNING

EMMANUEL HENRY*, DENIS MARCOTTE**, and MICHAEL SAMIS*
**AMEC, 2020 Winston Park Drive, Suite 700,*
Oakville, Ontario, L6H 6X7
***Department of Mineral Engineering, Ecole Polytechnique de Montréal, Quebec, H3T 1J4*

Abstract. Mine valuation under market and geological uncertainty is an active research area. Twenty years ago, a seminal paper by Brennan and Schwartz described the application of Real Option Theory to the valuation of mines where metal prices are volatile. The study focused mainly on the impact of metal price uncertainty on the value of a mine. Geological uncertainty was not considered. For a simple mine model, this paper describes the close analogy between the decision to process a mining block at a given date and the European call financial option. The value of the European call depends primarily on the share price model, the present share price, the price volatility and the time to expiry. A mining block is either processed when the metal price covers the processing costs or otherwise stockpiled as waste. Metal prices and technical variables like grades, recovery, and costs are uncertain. Using geostatistical simulations, the study shows that grade uncertainty may introduce asymmetries in the block value greater than metal price uncertainty. The asymmetries are more pronounced for blocks with larger uncertainty. Greater value is given presently to these blocks assuming the block grades are perfectly known at the time of mining. The extension of this concept from individual blocks to the mine scale is done by considering a mine panel as equivalent to a portfolio of European call options. Implications for strategic planning are illustrated with a gold mine panel-scheduling example. Gold price was modelled with a Geometric Brownian Motion process. The case study shows that the value of the panel and its development strategy depend on the level of geological uncertainty and price volatility. However, the example shows that the benefits of optimising the panel under geological uncertainty is an order of magnitude below the benefits of resolving the geological uncertainty.

1 Introduction

Mine valuation under market and geological uncertainty is an active research area. Twenty years ago, a seminal paper by Brennan and Schwartz (1985) described the application of Real Option Theory to the valuation of mines where metal prices are volatile. This theory relies on former developments in Finance Theory on the evaluation of financial derivatives (financial options).

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 501-510.

In the financial markets, an option is a contract giving the right, without the obligation, to buy or sell a share, or any financial instrument, at an agreed price and either at or before an agreed time in the future. An option that gives the right to buy a share at an agreed price (**E**) and a specific time in the future (**T**) is commonly called a European call option. **E** and **T** are called the exercise price and the time to expiry of the option, respectively. The share price (**S**) is usually volatile and at a time **T**, the option value ($\mathbf{v}_{\mathbf{opt}}$) will be:

$$\mathbf{v}_{\mathbf{opt}} = \max(\mathbf{S}_{\mathbf{T}} - \mathbf{E}, 0) \qquad [1]$$

If at time **T**, the share price is higher than the exercise price, then the owner of the option will be able to buy a share at price **E** and sell it immediately at price $\mathbf{S}_{\mathbf{T}}$, thus realizing a gain of $\mathbf{S}_{\mathbf{T}}$-**E**. If at time **T**, the share price is lower than the exercise price, the owner of the option will simply not exercise her option.

The value of an option, i.e. the price someone is ready to pay now to acquire the option (contract), depends on the price model and volatility, the current price, the exercise price, and the time to expiry. It is directly linked to the probability of the price being higher than the exercise price at expiry. Intuitively, the larger the time to expiry, the larger the price volatility, or the higher the current price, the higher the option value. Black and Scholes (1973), and Merton (1973), developed the first quantitative model for valuing European-like options with a share price following a Geometric Brownian Motion (Random Walk). Fig. 1 illustrates the value of a European call option as a function of the present price, a Geometric Brownian Motion price model, a time to expiry **T** of 1 year, and four annual volatilities (σ): 0%, 10%, 20%, and 30%. The exercise price **E** is $1. Fig. 1 shows that the value of the option increases with price volatility, and that the option can have a positive value even if the present price is below the exercise price. Practical valuation methods and algorithms for financial derivatives (a broader name for options) are found in Wilmott *et al.* (1995).

The Real Option Theory is the use of the Financial Option Theory to value real investments (see for details: Dixit and Pindyck, 1994, Amram and Kulatilaka, 1999, or Trigeorgis, 2000). Until recently, real option applications to mineral investments have considered mineral price volatility as the main source of uncertainty, and attempts to integrate geological uncertainty (as early as in Brennan and Schwartz, 1985, but see also Cortazar *et al.,* 2001, e.g.) were far from realistic. Carvalho *et al.* (2000) introduced a geostatistical simulation- and option pricing-based methodology to integrate geological models in the mine evaluation process. This paper focuses on the application of the Option Theory at the smallest scale in mines: the mining blocks. It illustrates the strong analogy between a mining block and a European call option, and shows how to use this analogy to value geological uncertainty and how geological and mineral price uncertainty interact. Implications for strategic planning are illustrated on a gold mine panel.

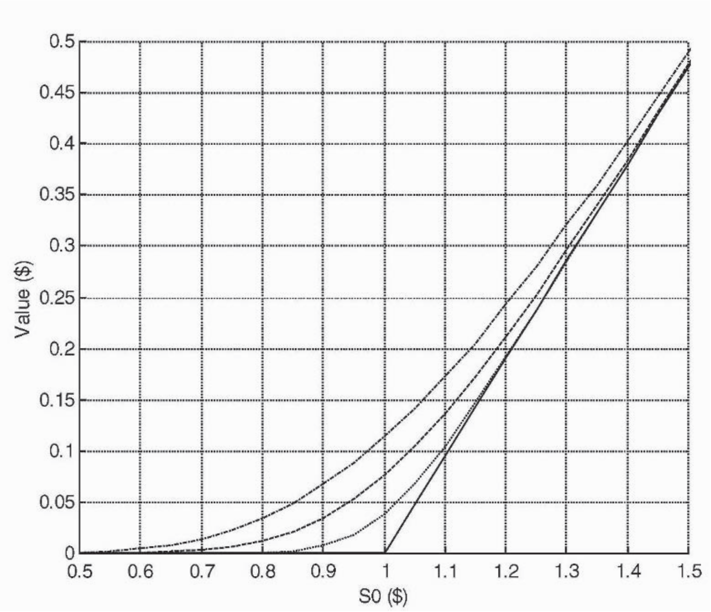


Figure 1. Value of a European Call Option as a Function of Current Price, S_0, and Price Volatility, $\boldsymbol{\sigma}$; **T** = 1 year, **E** = \$1, and $\boldsymbol{\sigma}$ = 0 % (continuous line), 10% (dotted line), 20% (dashed line), and 30% (dash-dotted line).

2 Analogy between a Mining Block and a European Call Option

A mining block contains **ton** tonnes of material at grade **g** of a mineral commodity sold at a price per unit **S**. The block may be developed or not. If the block is developed, it will be mined at a cost per tonne **m**, and then, either stockpiled as waste at a cost per tonne **stkp**, or processed at a cost per tonne **h** and marketed at a cost per unit **k**, depending on the benefit made by processing the block. Developing the block requires making the investment **dev** now, in order to be able to mine the block at time **T** from now. Recovery is denoted by **ρ**. Time discounting is ignored in this simple analogy, and the block is studied in isolation of the other blocks.

Traditionally, mine planners will make the decision to develop this block based on the block value $\mathbf{v}_{\mathbf{bl}}$:

$$\mathbf{v}_{\mathbf{bl}} = \mathbf{ton} \cdot \mathbf{g} \cdot \boldsymbol{\rho} \cdot (\mathbf{S} - \mathbf{k}) - \mathbf{ton} \cdot (\mathbf{m} + \mathbf{h}) \qquad [2]$$

and assuming grade and price are perfectly certain. If $\mathbf{v}_{\mathbf{bl}} > \mathbf{dev}$, the investment is worth making and the block will be mined, otherwise, the block will be left un-mined.

However, if **g** and/or $\mathbf{S} = \mathbf{S}_{\mathbf{T}}$ are uncertain at the time of decision, but if the true grade is known with certainty at the time of mining (for example, assuming that selection made on blast-hole sampling is exact, which is only a gross approximation), payoff from the block is similar to that of a European call option, with value:

$$\mathbf{v}_{\mathbf{bl}} = \max(\mathbf{ton} \cdot \mathbf{g} \cdot \boldsymbol{\rho} \cdot (\mathbf{S} - \mathbf{k}) - \mathbf{ton} \cdot (\mathbf{m} + \mathbf{h}), -\mathbf{ton} \cdot (\mathbf{m} + \mathbf{stkp})) \qquad [3]$$

Equation 3 has the same form as Equation 1, and $\mathbf{v}_{\mathbf{bl}}$ should therefore exhibit the same characteristics as $\mathbf{v}_{\mathbf{opt}}$:

- Grade uncertainty, like the mineral price, tends to increase the block value.
- Mineral price volatility tends to increase the block value.
- Grade uncertainty amplifies the effect of price volatility.
- The value of a block under price (and grade) uncertainty may increase with time **T**, if the volatility is large enough to compensate for the time discounting.

Table 1 summarizes the analogy between the mining block and a European call option.

Parameter	European Call Option	Mining Block
Time to Expiry	**T**	**T**
Exercise Price	**E**	**ton** (**m** + **h**)
Price	**S**	**ton g ρ** (**S** – **k**)
Cost of Not Exercising the Option	**0**	**ton** (**m** + **stkp**)

Table 1. Comparison between European Call Option and Mining Block.

3 Effect of Grade Uncertainty on the Value of a Mining Block

The European call option analogy of a mining block was investigated on a large panel in a gold mine, which is comparable to a portfolio of European call options.

The panel is made of 75 x 75 blocks, each of size 10 m x 10 m x 10 m. Geological information is provided by 50 m-spaced exploration drill holes. The gold distribution is lognormal with an average of 0.018 ounces per tonne (opt) and a coefficient of variation (CV) of 4. Gold grades are spatially variable with a strong nugget effect (40 % of the grade normal score variance, and ranges of 100 m north-south and 60 m east-west).

The mine planner is asked to determine which blocks should be developed now for production in two-years. Production and economic parameters are shown in Table 2.

The gold price is modelled with a Geometric Brownian Motion process following the equation:

$$\frac{\mathbf{dS}}{\mathbf{S}} = \boldsymbol{\mu} \cdot \mathrm{d}\mathbf{T} + \boldsymbol{\sigma} \cdot \mathrm{d}\mathbf{z} \qquad [4]$$

where μ is a trend, and d**z** is an increment to a standard Gauss-Wiener process. This model assumes no reversion to a long-term average, which is reasonable for gold, but not for most mineral commodities (see for example Schwartz, 1997, for more sophisticated commodity price models).

The focus of the first step is grade uncertainty while the price volatility and the effect of time are ignored. The panel is simulated 100 times on a fine grid and re-blocked to the nominal block size of 10 m x 10 m x 10 m. The simulations are then averaged to provide a map, illustrated in Fig. 2-a, similar to a kriged map. Mine planners are usually well aware that estimated (kriged) grade maps are uncertain, but for practical reasons, handle them as if they were certain, i.e. assuming they are an exact representation of the true grades, and overlooking local uncertainty associated with each block. Estimated grade maps are also generally smoother than in reality. Fig. 2-b shows a simulation outcome more representative of the true grade continuity.

Equation 3 is applied directly to the "certain" averaged model at time **T** = 0, assuming price is also certain and constant at $\mathbf{S_0}$ = \$350 per ounce. The blocks with value greater than the development cost **dev** should be developed and mined. Fig. 3-a shows the development outlines.

Recognizing that block grade-estimates are uncertain, and applying Equation 4 directly to each grade simulation outcome before averaging, gives the outlines in Fig. 3-b. The candidate area for development in Fig. 3-b is significantly larger than in Fig. 3-a, indicating the uncertain model generates more marginal blocks, i.e. blocks slightly higher than the economic break-even.

Parameter	Value	Variable Name in Text
Mining Production		
Block Tonnage	3,000 t	**ton**
Recovery	100%	ρ
Development Cost	\$1 /t	**dev**
Mining Cost	\$1 /t	**m**
Stockpiling Cost	\$0 /t	**stkp**
Processing Cost	\$3 /t	**h**
Marketing Cost	\$0 /oz	**k**
Price Model		
Model	Geometric Brownian Motion	
Present Price	\$350 /oz	$\mathbf{S_0}$
Trend	0%	μ
Annual Volatility	12%	σ
Risk-Free Discount Rate	5%	
Planning Periods		
Time to Mining	2 years	**T**

Table 2. Production and Economic Parameters.

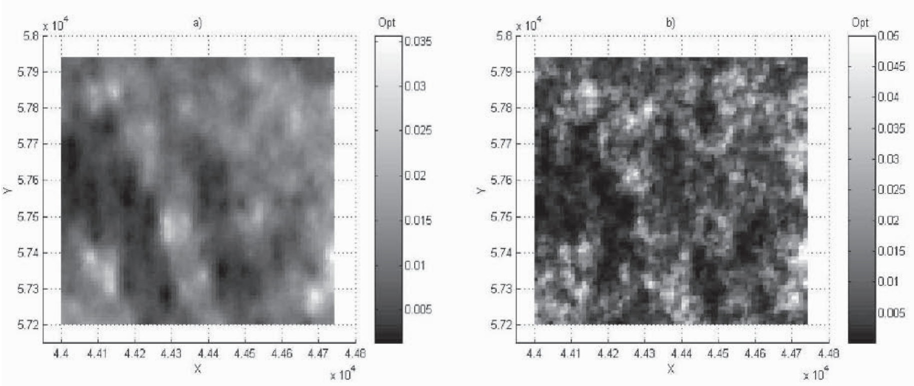


Figure 2. **a)** Average Grade of 100 Simulations; Reference Grade Model. b) Simulation Outcome Showing True Grade Variability.

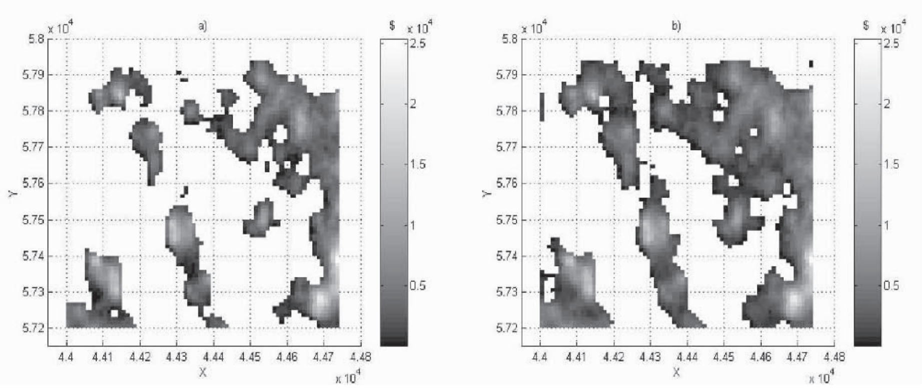


Figure 3. a) Value of the Blocks under Grade Certainty. b) Value of the Blocks under Grade Uncertainty. Only Positive Value Blocks are Shown.

Block values are plotted versus grades in Fig. 4. The black points correspond to the certain model, following the two linear equations:

$$\mathbf{v} = -\mathbf{ton} \cdot (\mathbf{m} + \mathbf{stk}) \qquad [5]$$

when $\mathbf{g} < (\mathbf{m} + \mathbf{h}) / \boldsymbol{\rho}\,(\mathbf{S} - \mathbf{k})$, and otherwise:

$$\mathbf{v} = \mathbf{ton} \cdot \mathbf{g} \cdot \boldsymbol{\rho} \cdot (\mathbf{S} - \mathbf{k}) - \mathbf{ton} \cdot (\mathbf{m} + \mathbf{h}) \qquad [6]$$

Light- and dark-grey points correspond to the uncertain model; light-grey points are for blocks with a CV greater than the average CV, 0.74, and dark-grey points, for blocks with a CV lower than 0.74. The figure shows that:

- Lower-grade blocks have more relative uncertainty (higher CVs) than higher-grade blocks. This is particular to this example and cannot be generalized.
- The higher grade-uncertainty, the higher the option value, as demonstrated by the light-grey points being above the dark-grey points.
- Grade uncertainty decreases the effective break-even cut-off grade (by approximately 20%, from the theoretical 0.011 opt down to 0.009 opt).

This result is in complete agreement with the findings made on financial options.

The impact of grade uncertainty is relatively large on the cut-off grade. However, this impact may be dampened, by the stockpiling cost for example. If **stkp** = \$0.5 per tonne instead of \$0 per tonne, grade uncertainty decreases the break-even cut-off grade by 10% only.

4 Cumulated Effect of Grade and Price Uncertainty on the Value of a Mining Block

Mineral price is usually considered as being volatile, i.e. uncertain, rather than certain. Both grade and price are then random variables that multiply each other in Equation 3. Price variance only is a function of time, with $\boldsymbol{\sigma_S(0)}^2 = 0$, and $\boldsymbol{\sigma_S(T)}^2$ increasing infinitely with time **T** for a Geometric Brownian Motion process.

In order to evaluate the impact of combining grade and price uncertainties, prices were simulated 1,000 times. No attempt was made to best fit the parameters to historic gold prices. However, the set of parameters in Table 2 is considered fair for the sake of the demonstration. Block values were calculated using Equation 3 and then averaged for each grade simulation and price simulation.

Fig. 5 shows block values as function of grade at **T** = 2 years (σ_S = 17%). In this example, grade uncertainty is far more important than price uncertainty.

5 Optimisation under Uncertainty

Optimisation of development outlines was performed on the maps shown in Fig. 3-a and 3-b using the assumption of uncertain grades. As commented earlier, uncertainty broadens somewhat the optimal design suggested by the certain value model. The broadened design increases the chances of capturing high-grade. The different designs suggested by the certain and uncertain models were applied successively to each of the 100 grade simulation scenarios. The average net value realized was then calculated. The procedure was also applied for the 100 optimal designs based on simulations, one for each simulation outcome. Each design was applied to all realizations and the average taken over the 100 x 100 possible combinations.

The results reported in Table 3 suggest that the most valuable mine design is the one which recognises the grade uncertainty (The comparison alone does not constitute a proof but provides useful indications).

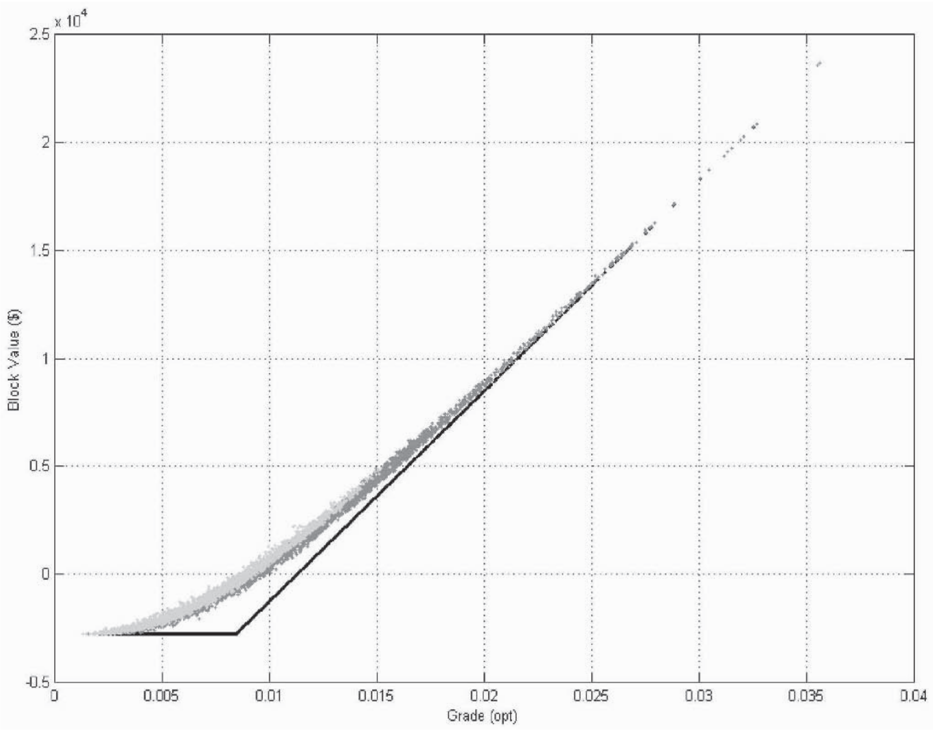


Figure 4. Block Value as a Function of Grades; Mining and Stockpiling Cost: $1 /t; Aligned Black Points: Grade Certainty; Dark-Grey Points: Grade Uncertainty with Block CV < 0.74; Light-Grey Points: Grade Uncertainty with Block CV > 0.74.

Design Based on	Base Scenario		Halved Variogram Range Scenario	
	Average Value	Relative Difference*	Average Value	Relative Difference*
Certain Grades	$4.5 M	0%	$2.4 M	0%
Uncertain Grades	$4.8 M	+6%	$2.7 M	+14%
Individual Simulations (Average)	$1.7 M	-61%	-$0.1 M	-103%

* To Certain Grade Model

Table 3. Value of Design Alternatives if Panel Mined at **T** = 2 years.

The design obtained by recognising grade uncertainty improves the certainty-based design value by 6% only. This is a small improvement and other biases in the mine optimisation parameters, such as assay results or cost estimates, would likely affect the design in the same order of magnitude as grade uncertainty. Table 3 also highlights that simulations are not useful in isolation: The best result achieved on a single simulation is about one-third of the result achieved using the option-based approach.

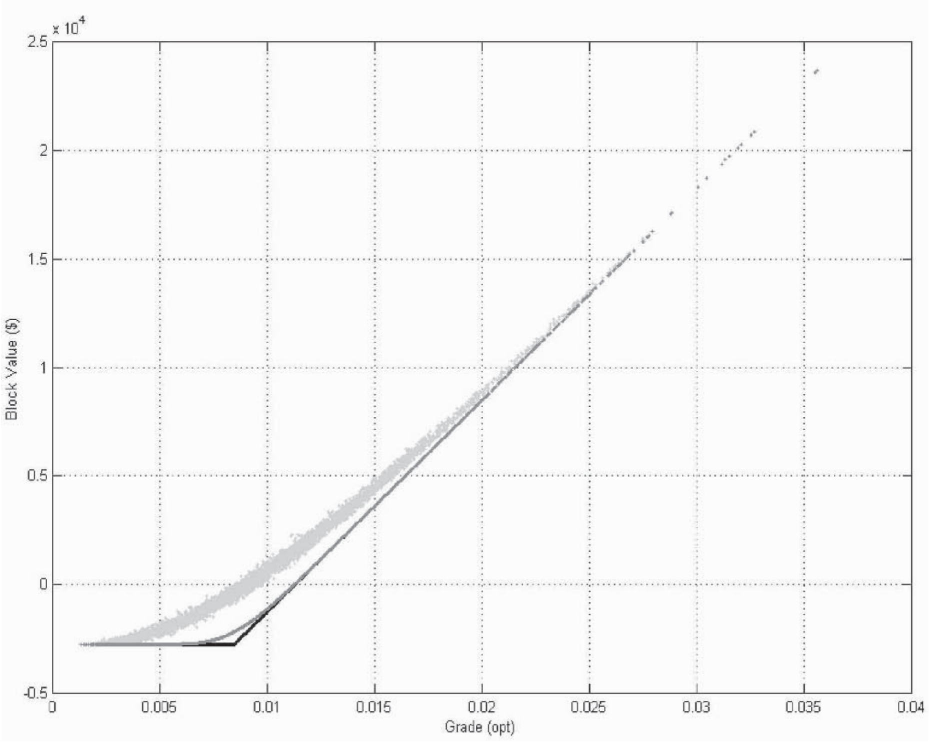


Figure 5. Block Value as a Function of Grade at **T** = 2 years; Mining and Stockpiling Cost: \$1 /t; Aligned Black Points: Value with Grade and Price Certainty; Dark-Grey Points: Value with Grade Certainty and Price Uncertainty (σ_S = 17%); Light-Grey Points: Value with Grade and Price Uncertainty.

The study was repeated assuming a nugget effect of the normal scores of the grades equal to 10% of the sill, and halved ranges (50 m north-south and 30 m east-west). Lower panel values were obtained, as shown in Table 3, but the uncertainty-based optimisation is now 14% higher than the certainty-based optimisation.

6 Discussion and Conclusion

The methodology described in this paper provides a framework for integrating all sources of uncertainty, technical and/or financial. The complexity of Equation 3 can (and will most probably) be increased to include other important aspects of project evaluation, such as multiple minerals or foreign exchange uncertainty.

Grade uncertainty, as any other technical uncertainty, is project-specific and may or may not be discounted for risk, depending on the project analyst's application of Finance Theory. This is not the same as ignoring project uncertainty. Some analysts may view geological uncertainty as project specific and diversifiable, so that a risk adjustment is not necessary. Others may be of the opinion that geological uncertainty cannot be mitigated through diversification and would consider applying an appropriate risk adjustment. Systemic uncertainty in mineral prices, however, is not diversifiable, so prices are risk-adjusted. This was realised practically by performing risk-neutral price simulations instead of "real" price simulations.

The methodology is especially interesting for economically marginal projects (or project areas) only. It may be useful for evaluating near end-of-life investments, or capital-intensive push-backs in large open-pits. Uncertainty, technical or financial, is not that relevant for clearly uneconomic or clearly economic projects.

The impact of recovery was not studied in this paper. Recovery less than 1 will decrease the value of the block in Equation 3. It will also decrease the option value generated by grade uncertainty, by dampening the grade standard deviation.

The interest of the mining industry for uncertainty-based optimisation is encouraging. However, it is important to stress that the value added by an uncertainty-based optimisation may be an order of magnitude less than the value lost by not resolving the uncertainty. In the panel example illustrated here, optimisation under grade uncertainty improves the project value by 6%, for a value of $4.8 M. In comparison, the panel value if the true grade was known with certainty would be $15.6 M in average. In other words, geological uncertainty adds value to individual blocks, but destroys two-third of the true potential project value.

References

Amram, M. and Kulatilaka, N., *Real Options, Managing Strategic Investments in an Uncertain World*, Harvard Business School Press, 1999.

Black, F. and Scholes, M., The Pricing of Options and Corporate Liabilities, *Journal of Political Economics* vol. 81, 1973, p. 637-659.

Brennan, M. and Schwartz, E., Evaluating Natural Resource Investments, *Journal of Business*, vol. 58, no. 2, 1985, p. 135-157.

Carvalho, R., Remacre, A., and Suslick, S., Geostatistical Simulation and Option Pricing Techniques: A Methodology to Integrate Geological Models in the Mining Evaluation Projects, *6th International Geostatistical Congress*, vol. 1, 2000, p. 1-10.

Cortazar, G., Schwartz, E. and Cassasus, J., Optimal Exploration Investments under Price and Geological-Technical Uncertainty : A Real Options Model, R&D Management, vol. 31, no. 2, 2001, p. 181-189.

Dixit, A. and Pindyck, R., *Investment under Uncertainty*, Princeton University Press, 1994.

Merton, R., Theory of Rational Option Pricing, *Bell Journal of Economics and Management Science*, vol. 4, no. 1, 1973, p. 141-183.

Schwartz, E., The Stochastic Behavior of Commodity Prices: Implications for Valuation and Hedging, *Journal of Finance*, vol. LII, no. 3, 1997, p. 923-973.

Trigeorgis, L., *Real Options, Managerial Flexibility and Strategy in Resource Allocation*, MIT Press, 2000.

Wilmott, P., Howison, S. and Dewyne, J., *The Mathematics of Financial Derivatives*, Cambridge University Press, 1995.

CLASSIFICATION OF MINING RESERVES USING DIRECT SEQUENTIAL SIMULATION

AMILCAR SOARES [1]

[1] *Environmental Group of the Centre for Modelling Petroleum Reservoirs, CMRP/IST, Av. Rovisco Pais, 1049-001 Lisbon, Portugal.*
e-mail: ncmrp@alfa.ist.utl.pt

Abstract

In mining operations, ore types are usually defined on the basis of technological criteria such as mining costs, processing plant performance and commercial costs, among others. Ore-type classification based on cut-off grades of estimated feed grades, tend to be biased when metal values or costs of ore type treatment are not linearly dependent on the feed grades. This paper presents an ore-type classification methodology based on jointly simulated grades. Direct sequential simulation (dss) and co-simulations (dscs) are the simulation techniques proposed to generate equiprobable images of different metal grades. Metal values and operating costs are then computed with several simulated grades of a block, in order to *a priori* classify the block, assigning it to the ore type which maximizes the profit or minimizes the costs of misclassification.

A case study on the Neves Corvo mine illustrates the proposed methodology.

1 Introduction

In most mining operations, ore types are defined on the basis of technological criteria such as mining costs, processing plant performance and commercial costs, among others. Cut-off values of feed grades, together with geological criteria, are normally used for *a priori* classification of mining reserves into different ore types. However this classification can be severely biased when metal values or costs of ore type treatment are not linearly dependent on the feed grades and the *a priori* classification of mining reserves is performed on the estimated grades of blocks and stopes.

Suppose the value of a given stope is not a linear function of its grades, for example if metal recovery is highly non-linearly related with the feed grade, or the commercial costs have a non-linear dependence on penalty grades. Then, the decision of sending that stope to a given ore type stockpile based on the estimated feed grades cannot be the one that maximizes the profit of the stope.

If one knows not only the mean grade of a stope or block but also the local cumulative distribution function (cdf), the idea of the proposed methodology is to apply known non-linear functions of metal recovery, values, costs etc., to the cdf of a given stope, rather than to its estimated mean grade, in order to choose the best ore type stockpile.

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 511-522.

Direct sequential simulation (dss) and co-simulations (dscs) are the simulation techniques proposed to generate equiprobable images of different metal grades. Metal values and operating costs are then computed with several simulated grades of a block in order to *a priori* classify the block in the ore type, which maximizes the profit or minimizes the costs of misclassification.
A case study on the Neves Corvo mine will illustrate the proposed methodology.

2 Case study: Neves Corvo mine

Neves Corvo is an underground tin-copper mine, which has been producing since the end of 1988. Considering the existing three orebodies, this study focuses on the Graça orebody that has been mined by a highly selective mining method (drift & fill) to maximize ore-type classification.
The main economic metal present in the ore is copper and the mineralisation can be described as being of the fissural or stockwork type. It is composed of veinlets and strings of sulphides and quartz, which cut mainly acid volcanic rocks, concordantly or not with the schistosity. The sulphides are mainly pyrite and chalcopyrite and the thickness of the veinlets may vary from a few millimetres to a few decimetres. The spatial distribution of the veinlets is highly irregular – as well as that of the grades – and does not show, in most situations, to be controlled by any particular geological feature. Cassiterite and stannite (tin and copper sulphide) are the main tin ores.
Two main ore types are defined in the Graça orebody: cupriferous ore (MC) and tin ore (MS), which are treated in different plants. The MS plant recovers copper and tin while the MC plant recovers only copper.
Data of Cu and Sn coming from drill-hole samples are available for this study.

3 Direct Sequential Simulation and Co-simulation

The principle of direct sequential simulation (dss) can be summarized as follows:
If the local cdfs are centred at the simple kriging estimate

$$z(x_u)^* - m = \sum_{\alpha} \lambda_\alpha(x_u)\left(z(x_\alpha) - m\right)$$

with a conditional variance identified by the simple kriging variance $\sigma^2_{sk}(x_u)$, the spatial covariance model or semivariogram is reproduced in the final simulated maps (Journel, 1994). The problem is that this simulation approach does not reproduce the histograms of the original variables (the local cdf cannot be fully characterized by only the local mean and variance).
The idea of direct sequential simulation (Soares, 2001) is to use the estimated local mean and variance, not to define the local cdf but to sample the constant global cdf $F_Z(z)$. Intervals of z are chosen from $F_Z(z)$, and simulated values $z^s(x_u)$ are subsequently sampled from them. These intervals are "centred" at the simple kriging estimate $z(x_u)^*$, being the interval range dependent on the simple kriging estimation variance $\sigma^2_{sk}(x_u)$ (Soares, 2001).

One of the main advantages of the proposed dss algorithm over traditional sequential indicator simulation (sis) and sequential Gaussian simulation (sGs) to simulate continuous variables is that it accommodates joint simulation of original variables without any prior indicator or Gaussian transformation.

In this case, the joint simulation of both metals Cu and Zn follows the Bayes rule. That means, that the simulation of a pair of values from a bi-variate distribution, say $F(Z_1,Z_2)$, is equivalent to generating the first value z_1 from the marginal distributions $F(Z_1)$ and the second from the conditional distribution, $F(Z_2|Z_1{=}z_1)$. In a spatial process with two correlated variables, $Z_1(x)$ and $Z_2(x)$, the first value z_1 is simulated from $F_{Z1}(x_u; z) = \text{prob}(Z_1(x_u)<z)$ at the location x_u and, afterwards, z_2 is generated from the conditional distribution $\text{prob}(Z_2(x_u)<z \mid Z_1(x_u) = z_1)$ (Almeida and Journel, 1994). The first variable is simulated with direct sequential simulation and the second variable using direct sequential co-simulation (Soares, 2001).

The same algorithm is then applied to simulate $Z_2(x)$ assuming the previously simulated $Z_1(x)$ as the abundant (known at every node) secondary variable. Co-located simple co-kriging is used to calculate $z_2(x_u)^*$ and to estimate $\sigma^2_{sk}(x_u)$ conditioned to neighbourhood data $z_2(x_\alpha)$ and the co-located datum $z_1(x_u)$ (Goovaerts, 1997).

One crucial issue of this sequential approach regards the choice of the variable to be simulated first. In sequential simulation algorithms local conditional distributions are estimated with some approximations, for example, the conditioning data is limited to a subset of samples (Gomez-Hernandez and Journel, 1993); hence, for variables with different spatial continuity patterns, the result is not independent of the order of the chosen sequence of variables to be simulated. Hence, practical criteria regarding the spatial pattern of the variables and its relative importance in the physical phenomenon are normally applied. In this case, Cu grades are simulated first – through direct sequential simulation – since Cu is the main metal, the most valuable one and, on top of that, reveals a more continuous spatial pattern.

4 Classification of mining reserves in ore types.

The usual procedure of classification of mining reserves in ore types consists in using the estimated average grades of each block as a threshold criterion to classify it. If a block value is a non-linear function of its grades, classification can be severely biased when performed with estimated average grades.

The idea of the proposed classification can be summarized in two basic points:

The classification is based on simulated grades, rather than estimated ones, which allow preserving the histograms of different metals, spatial pattern continuity and the spatial relationship between them.

The criteria to classify one given block in ore types will be based on the maximization of a profit function, or minimization of a cost function, applied to the joint simulated values of the block.

In the case study of the Neves Corvo mine, the processing plants have different metal recoveries and costs. The commercial costs, which include transport, shipping, insurance, treatment and refinement charges are also different for both metals, copper and tin.

Hence, the value of a given stope can be viewed as the difference between the metal value minus the treatment and the commercial costs. Suppose a block is located at x_u

with the feed grades of Cu and Sn: $z(x_u)$ and $y(x_u)$. The value of one tonnage of a block of MS ore type can be summarized as:

$$v_{MS}\,(x_u)= [s_{Cu}\ \text{-}c_{Cu.}/z_c].z\,(x_u).\ \ \eta_{Cu} + [s_{Sn}\ \text{-}c_{Sn}/y_c].y\,(x_u).\ \ \eta_{Sn} - mc\text{-}plc_{MS} \qquad [1]$$

that is, the sum of the copper value (net value minus costs) plus the tin value minus the mining costs and the MS plant costs. The tonnage of a block value of MC:

$$v_{MC}(x_u)= [s_{Sn\cdot}\ \text{-}c_{Sn}/y_c].y(x_u).\ \ \eta'_{Cu} - mc\text{-}plc_{MC} \qquad [2]$$

which is the sum of the copper value minus the mining costs and the MC plant costs.

with: s – metals price of Sn and Cu ; $\eta_{Cu},\ \eta_{Sn}$ – Cu and Sn recovery at MC plant; η'_{Cu} – Cu recovery at MS plant; $z_c\,,\, y_c$ – concentration grades; $c_{Cu}\,,\ c_{Sn}$ – commercial costs; *mc* – mining costs; *plc* –plant costs

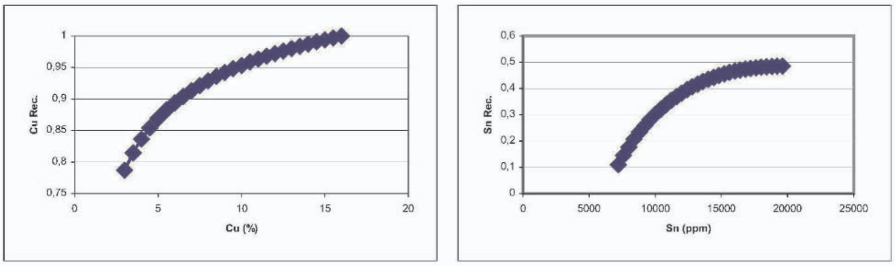


Figure 1. Metal recovery of Cu vs Cu (%) at MC plant and metal recovery of Sn vs Sn at MS plant.

Metal recovery of Cu and Sn are non-linear functions of the feed grades. Figure 1 shows the metal recovery of Cu *vs* Cu (%) at the MC plant and metal recovery of Sn *vs* Sn (%) at the MS plant.

The criterion to classify x_u as MC or MS is the maximization of the profit, $v_{MS}{}^{l}\,(x_u)$ or $v_{MC}{}^{l}\,(x_u)$, for the entire set of realizations $l = 1, Ns$, or, in other words, the minimization of the costs of misclassification. That is, x_u will be classified as the ore type that maximizes the value for the entire set of realizations. x_u is classified as cupriferous ore if:

$$\frac{1}{N_s}\sum_{l=1}^{Ns} v_{MC}{}^{l}\left(x_u\right) > \frac{1}{N_s}\sum_{l=1}^{Ns} v_{MS}{}^{l}\left(x_u\right) \qquad [\,3\,]$$

x_u is considered as tin ore otherwise.

Note that as the value v is a non-linear function φ of the feeding grades z, $v^{l}(x_u) = \varphi[z^{l}(x_u)]$, a different result is achieved when this criterion is applied to an average grade of Cu or Sn at x_u:

$$\sum_{l=1}^{Ns}\varphi\left(z^{l}\left(x_u\right)\right) \neq \varphi\left(\sum_{l=1}^{Ns} z^{l}\left(x_u\right)\right)$$

An alternative criterion to [3] could be chosen in terms of costs rather than profits: the minimization of the costs of misclassification (Goovaerts, 1997). Suppose the following loss functions: the loss associated with classifying x_u as MC

$$L_1^{\,l}(x_u) = \begin{cases} 0 & \text{if } v_{MC}^{\,l}(x_u) > v_{MS}^{\,l}(x_u) \\ v_{MS}^{\,l}(x_u) - v_{MC}^{\,l}(x_u) & \text{otherwise} \end{cases}$$

and the equivalent loss associated with classifying x_u as MS:

$$L_2^{\,l}(x_u) = \begin{cases} 0 & \text{if } v_{MS}^{\,l}(x_u) > v_{MC}^{\,l}(x_u) \\ v_{MC}^{\,l}(x_u) - v_{MS}^{\,l}(x_u) & \text{otherwise} \end{cases}$$

The N simulated images allow calculating the average loss attached to the two types of classification:

$$\varphi_1(x_u) = \sum_{l=1}^{Ns} L_1^{\,l}(x_u) \text{ and } \varphi_2(x_u) = \sum_{l=1}^{Ns} L_2^{\,l}(x_u) \qquad [\,4\,]$$

the location x_u is declared to belong to MC or MS if it minimizes the corresponding average losses:

$$\varphi_1(x_u) > \varphi_2(x_u)$$

meaning that the costs of classifying x_u as MC are greater than the costs of classifying x_u as MS, hence x_u is classified as MS;

$$\varphi_2(x_u) > \varphi_1(x_u)$$

meaning that the costs of classifying x_u as MS are greater than the costs of classifying x_u as MC, hence x_u is classified as MC.

Both approaches [3] and [4] are equivalent and give the same results which are presented and compared with results following the more traditional routine of ore-type classification based on estimated grades.

5 Results

5.1 DATA ANALYSIS

Histograms of Cu and Sn were calculated from 524 samples from boreholes of the Graça orebody (Figures 2a and b). The Cu/Sn bi-plot shows the relationship between both elements (Figure 3). Spatial continuity main patterns of Cu and Sn can be summarized in the following: both Cu and Sn present a similar isotropic behaviour, modelled by an exponential model. Sn variogram presents a clear “nugget effect”, representing 20% of the total variance, which is probably linked to the spatial dispersion of main tin mineralisations: cassiterite and stanite.

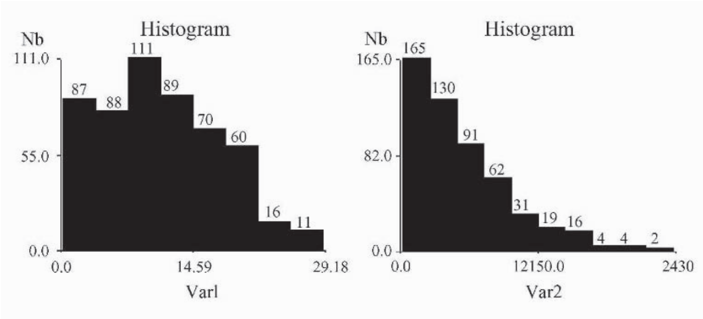
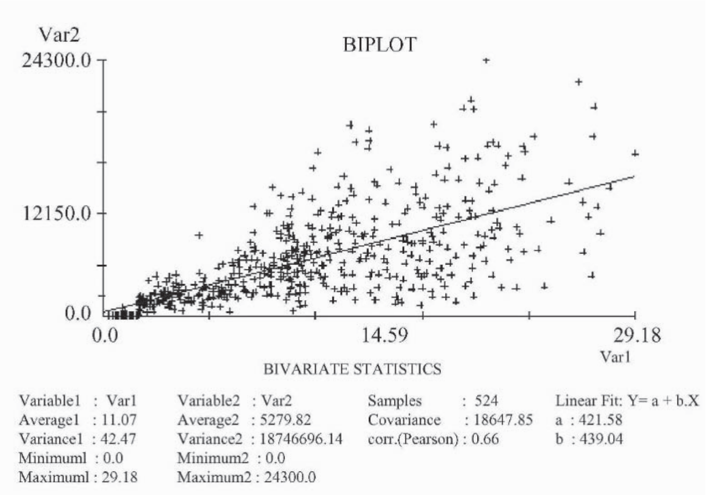


Figure 2. Histograms of Cu (a) and Sn (b).



```
Variable1  : Var1      Variable2  : Var2            Samples          : 524        Linear Fit: Y= a + b.X
Average1   : 11.07     Average2   : 5279.82         Covariance       : 18647.85   a : 421.58
Variance1  : 42.47     Variance2  : 18746696.14     corr.(Pearson)   : 0.66       b : 439.04
Minimum1   : 0.0       Minimum2   : 0.0
Maximum1   : 29.18     Maximum2   : 24300.0
```

Figure 3. Bi-plot of Cu/Sn

From the bi-plot of Figure 3 one can visualize two populations with different behaviours regarding the correlation between Cu and Sn. If the total set of samples is split by a Cu threshold of 10%, the values with a Cu content lower than 10% show a higher correlation coefficient (r = .71) (Figure 4a) than the values with a Cu content higher than 10%, which do not present a significant correlation with Sn (r = .37) (Figure 4b). Cu/Sn cross variograms computed for those populations confirm the distinct spatial co-regionalisation behaviours.

5.2 JOINT SIMULATION OF CU AND SN

Cu and Sn grades were simulated in a regular grid of points (200x 110 x 20 nodes of 1x1x1m.). Sn values were simulated using direct sequential co-simulation assuming previously simulated maps of Cu as a secondary variable.

Simple collocated co-kriging was used for the estimation of Sn at each node of the regular grid visited during the sequential procedure:

$$z_1(x_u)^* = \sum_{\alpha=1}^{N} \lambda_\alpha(x_u)(z_1(x_\alpha) - m_1) + \lambda_u(x_u)(z_2(x_u) - m_2) + m_1 \qquad [\,5\,]$$

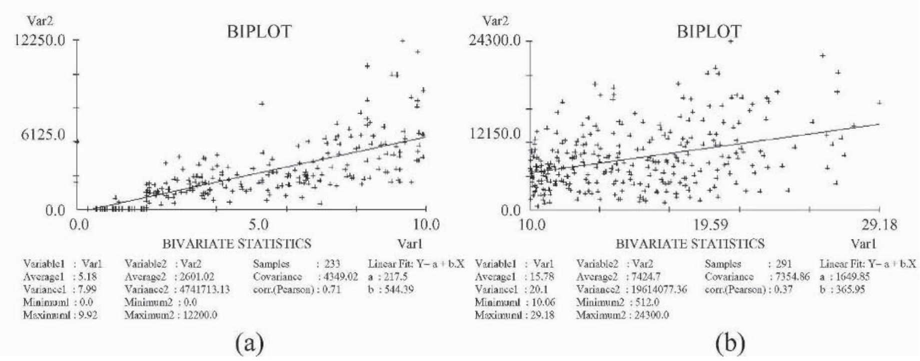


Figure 4. Bi-plot of Cu/Sn for the population with low grades of Cu a); and for the population with high Cu grades.

Collocated co-kriging is implemented with the Markov-type approximation of co-regionalization models: cross correlograms between *z1* and *z2*, $\rho_{z1,z2}(h)$ are determined by the correlation coefficient $\rho_{z1,z2}(0)$ and the correlogram of $_{z1}$ $\rho z1(h)$: $\rho_{z1,z2}(h)= \rho_{z1,z2}(0). \rho_{z1}(h)$ (Goovaerts, 1997). Two local co-regionalisation models between Cu and Sn (described in 5.1) were adopted with the Markov-type approximation: "low" grades of Cu (<10%) with a correlation coefficient r = .71 and "high" grades of Cu (≥10%) with r = .37. An example of level 1, with "low" and "high" grades of Cu is shown in Figure 5. Note that in this case, under the Markov-type approximation, to estimate a local Sn value at the location x_u, the co-regionalisation model is dictated by the correlation coefficient of x_u.

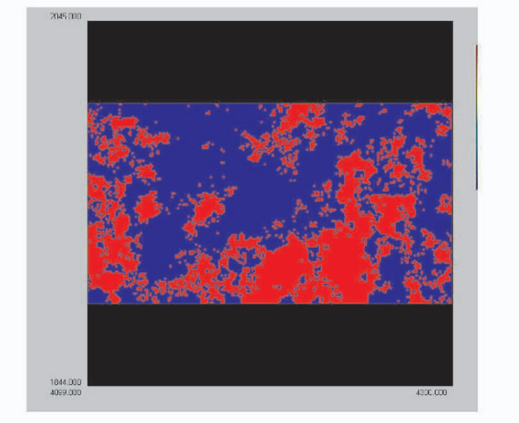


Figure 5. Estimated maps of "low" (red) and "high" grades (blue) of Cu

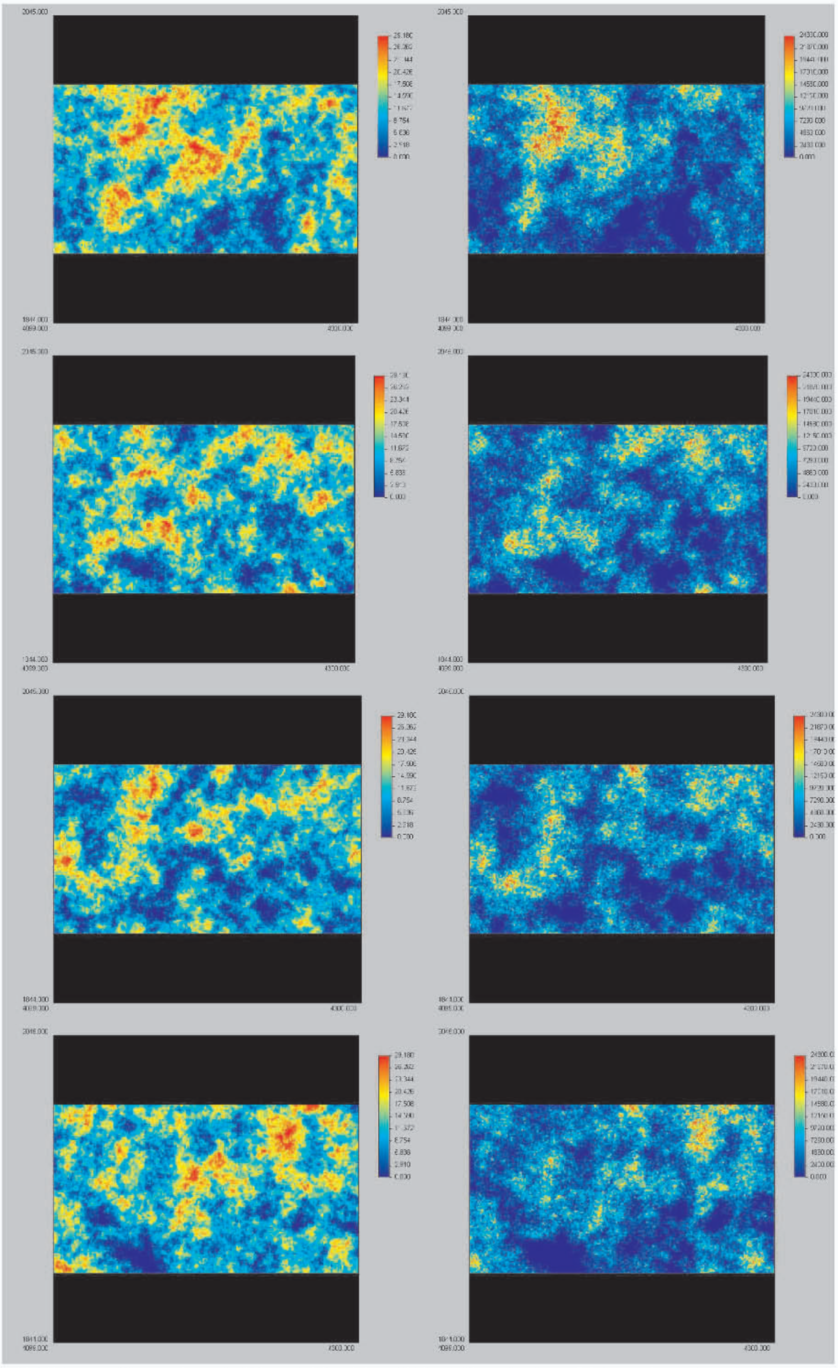


Figure 6. Direct Sequential Simulation and Co-Simulation : Four co-simulated pairs of Cu (left column) and Sn.

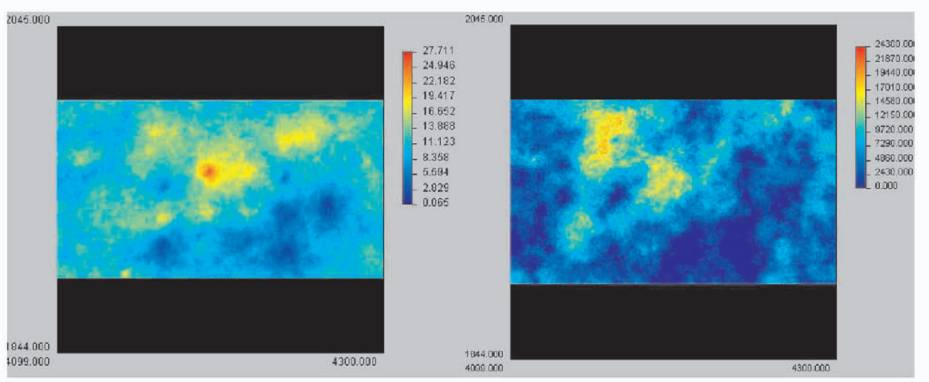


Figure 7. Average of 20 simulated maps of Cu (left) and Sn (right).

At the Cu “high” grades population (blue area of Figure 5), with a correlation coefficient r = .37, there is practically no influence of the secondary variable (simulated Cu grades). A set of 20 realizations of Cu and Sn were simulated for the entire area. Examples of four pairs of Cu and Sn images are presented in Figure 6. One example (level 1) of the average of 20 simulated maps of Cu and Sn are presented in Figures 7a and 7b, respectively. Notice that the influence of Cu at the Sn simulations is significant only at the “low” Cu grades area, where the correlation coefficient is high.

Marginal histograms and correlation coefficients of simulated Cu and Sn show a quite good match with the equivalent sample statistics.

Fig. 8 shows the variograms of the same two realizations of first level for Cu (left column) and Sn (right column). There is a very satisfactory match between the theoretical model (imposed to the simulations and co-simulations) and the experimental variograms of simulated values.

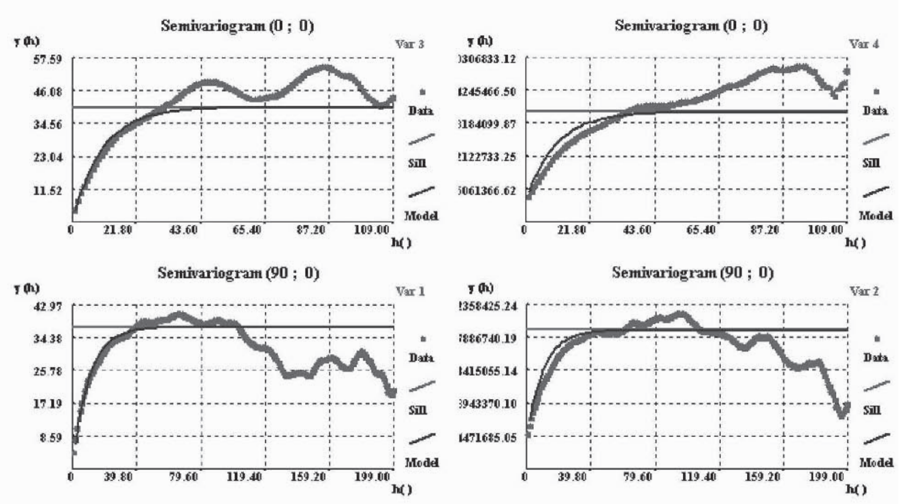


Figure 8. Variograms (experimental and model)of simulated values of Cu (left column) and Sn (right column) for two realizations.

5.3 CLASSIFICATION OF THE ORE TYPES

Each pair of simulated Cu and Sn values, at a given spatial location x_u, will feed the two profit functions [1] and [2] corresponding to the two different plant treatments and transport. Averaging out the profit of the 20 realizations for the two plant treatments will determine which ore type should be allocated to the spatial location x_u [3].

Sensitivity analysis has shown a high dependency of the profit of [1] on the tin metal prices. Four different tin metal prices were used, corresponding to those occurring during the period from the beginning of the mine's exploration until now. Figure 9 shows, in the right column, the two ore types classified on the basis of the simulated images for the four metal prices, from 4 US$/Lb (top) practiced in the end of the eighties, up to 2003 price of 2.4 US$/Lb (bottom). In the left column of Figure 9 the equivalent classification based on the average maps of Figures 7a and 7b is presented for comparison.

It is obvious that in both classifications the MS ore type decreases with the tin metal price. However, the classification of the average grade gives systematically higher proportions of MS than the simulations. This is quite expectable since the non-linear functions of Figures 1a and b, applied to a mean of a positively skewed histogram of Sn values (Figure 1b), tend to be greater than the mean of the non-linear transformation of each one of Sn grades. These highest proportions of MS reflect biased average-grades based classification: a systematic overestimation of MS proportion.

As a matter of fact, the continuous decreasing of tin metal price determined the very recent decision (taken in 2002) of the mine board to discommission the tin plant.

6. Final remarks

i) This paper presents the use of stochastic simulation images of different metal grades to classify mining reserves in ore types. When costs and values can be allocated to the main mining operations, classification of ore types based on joint simulated metals are a much more accurate and unbiased alternative than the classical procedure of classification based on estimated grades.

ii) This paper also shows that ore-type classification is a dynamic exercise of optimisation of future strategies, balancing historical decisions, the knowledge of reserves, and the near future of metal market prices, contracts, etc..

Considering the presented test case, when the decision of building a tin plant was taken, it was fully justified by the tin prices of that time. Once the tin plant was working, any classification should have been conditioned to its fixed and operational costs. According to the criteria followed in 5, most of the blocks are classified as MC. But that implies that a significant number of those blocks should remain unmined, given that the production capacity of the copper plant is limited. In this case, although we know that those blocks give, theoretically, more profit in the Cu plant, they should be sent to the tin plant as that will optimise the production capacity of both plants.

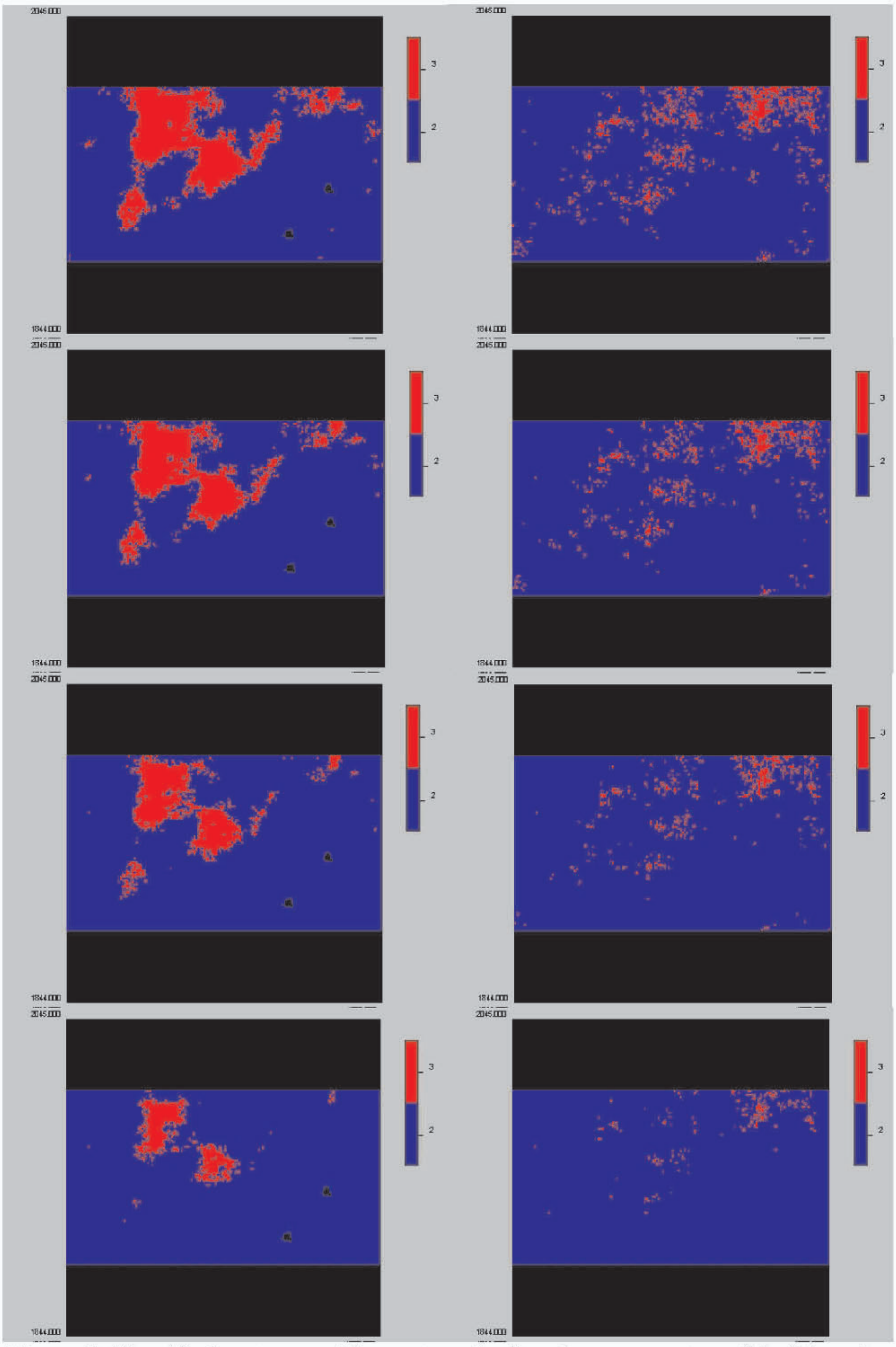


Figure 9. Classified oretypes – Tin oretype (red) and copper oretype (blue) based on 20 co-simulated pairs, and four different metal value of Tin (right hand side column); based on the average of the simulations (left hand side column).

iii) Finally, it is demonstrated that the combination of direct simulation and co-simulation is a very appropriate technique for the joint simulation of continuous variables. Recent applications of the dss can be found in environmental field , in soil pollution characterization (Franco C. et al, 2002), satellite image classification (Bio et al, 2002), ecological resources (Almeida et al, 2002) and in petroleum applications (Soares et al, 2001).

Acknowledgments

Somincor – Sociedade Mineira de Neves Corvo – is gratefully acknowledged for the permission to publish the data.
I would like to thank the valuable comments of two anonymous reviewers.

References

Almeida A., Journel A., 1994. Joint simulation of multiple variables with a Markov-type corregionalization model. Mathematical Geology 26(5): 565-588.

Almeida J., Bio A., Santos E., 2002. Use of Geostatistical Methods to Characterize Population and Recovery of Iberian Hare in Portugal. Proceedings of geoENV2002- Geostatistics for Environmental Applications. Barcelona.

Bio A., Carvalho J., Rosário L., 2002. Improving Satellite Image Forest Cover Classification with Field Data Using Direct Sequential Co –Simulation. Proceedings of geoENV2002- Geostatistics for Environmental Applications. Barcelona.

Caers J., 2000, Direct sequential indicator simulation. Proceedings of 6th International Geostatistics Congress. Cape Town. S.A..

Gomez-Hernandez, J., Journel A.G., 1993 - Joint Sequencial Simulation of MultiGaussian Fields. *Geostatistics TROIA´92*, Ed. Soares, A., Kluwer Pub., pp. 85-94.

Franco C.,Soares A. , Delgado J., 2002. Characterization of Environmental Hazard Maps of Metal Contamination In Guadiamar River Margins. Proceedings of geoENV2002- Geostatistics for Environmental Applications. Barcelona.

Goovaerts P., 1997. Geostatistics for Natural Resources characterization. Oxford University Press.pp 483.

Journel A.G., 1994. Modeling Uncertainty: Some Conceptual Thoughts. Geostatistics for the Next Century. ED Dimitrakopoulos R..kluwer Academic Pub.pp 30-43.

Soares A., 1998. Sequential Indicator Simulation with Correction for Local probabilities. Mathematical Geology, vol 30, N 6,, pp 761-765.

Soares A., 2001. Direct Sequential Simulation and Co-simulation. Mathematical Geology. 33-8. pp 911-926.

Soares A, Almeida J., Guerreiro L.. 2002. Incorporating Secondary Information using Direct Sequential Co-Simulation. To be published in Stochastic Modelling Vol II, American Association of Petroleum Geologists Pub.

USING UNFOLDING TO OBTAIN IMPROVED ESTIMATES IN THE MURRIN MURRIN NICKEL-COBALT LATERITE DEPOSIT IN WESTERN AUSTRALIA

MARK MURPHY[1], LYN BLOOM[2] AND UTE MUELLER[2]

[1] *Snowden Mining Industry Consultants, West Perth, Western Australia*
[2] *Edith Cowan University, Joondalup, Western Australia*

Abstract. Nickel and cobalt are key additives to modern alloys. The largest worldwide nickel-cobalt resources occur in surface laterite deposits that have formed during chemical weathering of ultramafic rocks at the Earth's surface. Geologically young deposits have formed by rapid weathering processes in tropical environments while older deposits that have formed in drier climates. At the Murrin Murrin mine in Western Australia the dry climate laterite deposits occur as laterally extensive, undulating blankets of mineralisation with strong vertical anisotropy and near normal nickel distributions. This deposit structure presents an estimation challenge for both classical and geostatistical resource estimation methods. In this paper, ordinary kriging and multiple indicator kriging estimation methods are applied to both the in situ and unfolded structural cases to obtain estimates for nickel and cobalt. Improvement in point grade estimation following the unfolding of the laterite blanket by vertical data translation prior to grade estimation is assessed in the light of close spaced grade control data. The results indicate that unfolding, particularly when combined with indicator kriging, improves both the nickel and cobalt estimates albeit only slightly in the case of cobalt.

1 Introduction

Nickel and cobalt are key metal additives in modern industry. Nickel and cobalt are primarily sourced from deep underground mines but the largest worldwide deposits where both metals occur are the near surface laterite deposits that have formed by weathering of ultramafic rocks in tropical or semiarid environments (Golightly 1981; Brand et al 1998). At Murrin Murrin in central Western Australia, surface weathering of ultramafic rocks in a semiarid environment has enriched nickel and cobalt to economically attractive concentrations approaching 2%Ni and 0.5%Co within smectite clay horizons. The nickel cobalt deposits at Murrin Murrin are flat lying, undulating blankets of 10 to 50 m thickness and lateral extents ranging from a few to tens of kilometres (Fazakerley and Monti, 1998).

2 The MM2 Dataset

One deposit area at Murrin Murrin, known as MM2 is the focus of this study. The data comprises samples collected from vertical drillholes during exploration and subsequent mining of the deposit. Exploration drilling was completed on a nominal 50 m square

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 523-528.

pattern and contains a local cluster of 12.5 m spaced holes (Figure 1, left). Grade control sampling was carried out on a 12.5 m square pattern to approximately 30 m below surface (Figure 1, right). For this study the drilling samples were accumulated into a composite length that matches the mining bench height of two metres. The exploration sampling was flagged as a subset of the grade control data and, both data sets were clipped to a boundary 30 m below surface and to a marginal ore processing threshold of combined nickel cobalt grade.

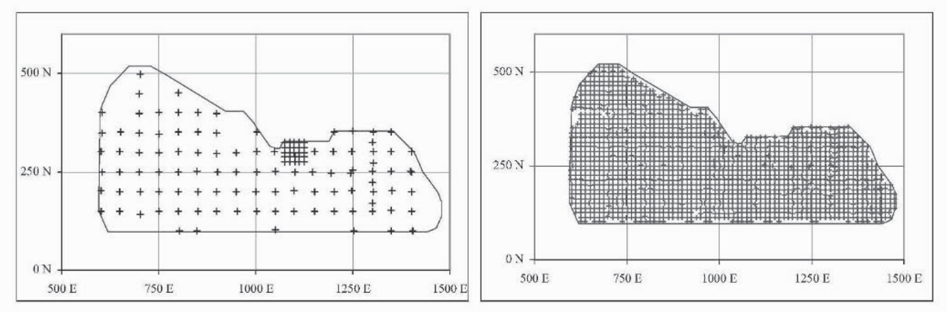


Figure 1. Exploration (left) and grade control (right) collar locations in the MM2 pit

For the purposes of this study the sampling from the grade control pattern is considered reality. Figure 3 shows cross sections through 250N (2:1 vertical exaggeration) with the 12.5 m spaced, bench height composites from grade control coded by nickel and cobalt grades within the ore envelope. These sections reveal that nickel forms a relatively continuous blanket of mineralisation with higher grades (>1.0 Ni%) defining an undulation in the nickel mineralisation across the area. In contrast, the high grade cobalt mineralisation (>0.06 Co%) is more pod-like but generally follows the blanket of nickel mineralisation.

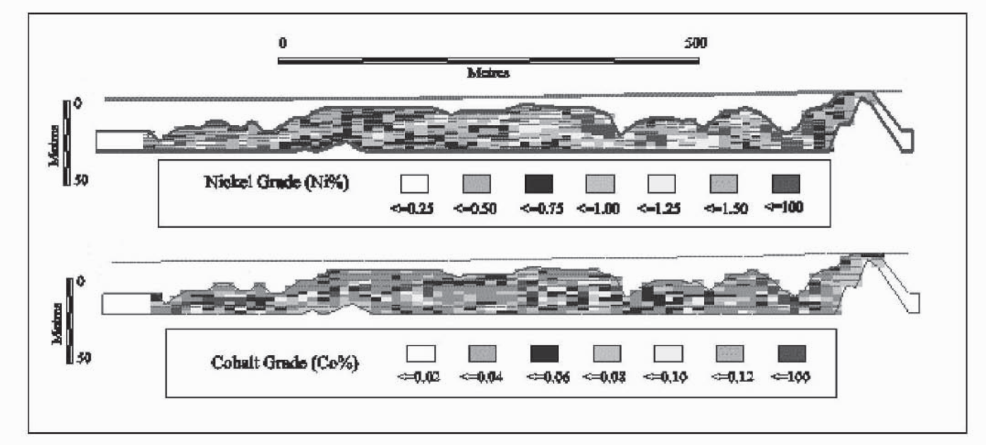


Figure 2. Cross section 250 N showing ore envelope and bench height composites

In Table 1 the summary statistics of both nickel and cobalt composites within the ore are compared for both the grade control and exploration sampling patterns. Declustered statistics were calculated using cell declustering to account for the clustered sampling in the exploration pattern.

Statistic	Nickel grade (%)			Cobalt grade (%)		
	Grade control	Exploration		Grade control	Exploration	
		Clustered	Declustered		Clustered	Declustered
Composites	13,414	1,046	1,046	13,411	1,046	1,046
Minimum	0.07	0.13	0.13	0.001	0.001	0.001
Maximum	2.67	2.23	2.23	0.887	0.887	0.887
Mean	0.85	0.84	0.80	0.058	0.054	0.053
Median	0.81	0.80	0.75	0.040	0.038	0.036
Standard deviation	0.38	0.39	0.38	0.054	0.054	0.056
CV	0.44	0.46	0.47	0.944	0.993	1.053

Table 1. MM2 grade summary statistics for grade control and exploration composites

The summary statistics show that the exploration sampling contains 1,046 samples compared to the 13,414 available from the final grade control pattern and that the nickel distribution is near normal while the cobalt distribution is highly skewed. Declustering produces in a minor reduction in the distribution means and a minor increase in data skewness. Accepting the grade control results as reality for this study, the exploration sampling statistics show that the exploration sampling pattern has been successful in determining the underlying mean and variability of both nickel and cobalt.

3 Unfolding

The large lateral extent and blanket geometry of nickel laterite deposits, combined with a strong vertical anisotropy, presents several problems for grade estimation from the exploration data. Of particular interest to mine planning is the correct reproduction of the lateral connectivity of higher grade zones as depicted in Figure 3. In Figure 3, a schematic cross section of a nickel laterite resource envelope and vertical drill holes is depicted against a backdrop of an estimation grid. The search neighbourhood used for estimation of the model nodes is shown as a flat lying ellipsoid with a shape dictated by the strong vertical anisotropy the deposit. A dashed line represents a surface of expected grade connectivity for this idealised deposit. It is assumed waste samples have been excluded from the estimation method.

In Figure 3, where drill holes are close together (near block A) or where the lies ore horizontally (near block B), the grade zones in the drill holes are reflected in the estimation model. However, where drilling is widely spaced and/or there is undulation of the surface of grade connectivity counter intuitive estimation results may occur (such as block C and block D). Problems of geometric controls arrecting grade estimation in situations of folded or undulated geometry have been recognised by prior authors (Wellmer & Giroux 1980, Dowd et al 1988, Lambert 2000, Sahin et al 1998, Sides and da Silva 1996). These authors have proposed several methods to remove estimation artefacts including domaining areas of similar geometry, data translation and application of local coordinate systems.

In this study, the vertical translation method was used to improve the grade connectivity of nickel grades within the study area (Murphy et.al. 2002).

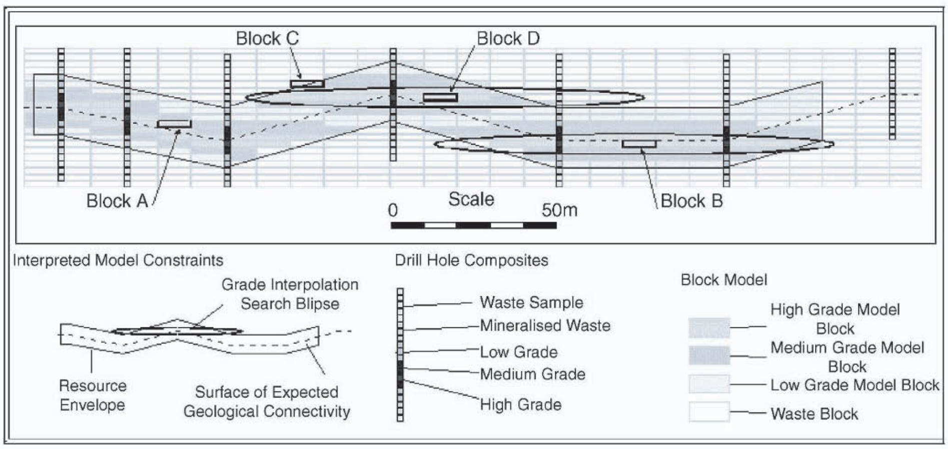


Figure 3. Schematic estimation model from vertical drilling in a finite domain

4 Variography

Traditional and indicator semivariograms were computed for the exploration data in both the in situ and unfolded data configurations. Twelve indicator thresholds were applied to both elements in 0.1%Ni increments ranging from 0.5 to 1.6%Ni, and 0.01%Co increments from 0.03 to 0.14%Co. The blanket geometry of the nickel mineralisation dictates that the minor axis of continuity is the downhole direction. Therefore, horizontal-plane semivariogram maps were used to test for the direction of maximum continuity in the horizontal plane and direction variograms were then computed for the axes of anisotropy.

For nickel, the traditional variography exhibits geometric anisotropy in the study area with a major axis of continuity as azimuth 70°. The variogram has a low nugget effect (0.02 of a sill of 1.00) and three nested spherical structures were fitted to the experimental data (0.45 sill, 7m x 30 m x 30m; 0.30 sill, 12 m x 50 m x 50 m; 0.23 sill, 15 m x 75 m x 100m). The variography of unfolded data has slightly longer ranges (0.45 sill, 7m x 30 m x 30m; 0.30 sill, 12 m x 60 m x 70 m; 0.23 sill, 15 m x 75 m x 200m). The nickel indicator semivariogram surfaces revealed patterns of rotational anisotropy with where the lower nickel thresholds having greater continuity east-west and higher thresholds having longer NE-SW continuity. There is a pattern of decreasing ranges and increasing nugget effect with increasing indicator nickel threshold and slightly longer ranges interpreted for the unfolded case.

For cobalt, the traditional semivariogram has a major axis azimuth of 100° and a nugget effect of 0.25. Again three nested structures were modelled for the in situ data (0.40 sill, 6 m x 20 m x 60m; 0.25 sill, 8 m x 50 m x 70 m; 0.23 sill, 10 m x 150 m x 200 m) and unfolded cases (0.40 sill, 8 m x 20 m x 20 m; 0.25 sill, 9 m x 30 m x 30 m; 0.23 sill, 10 m x 40 m x 40 m) with unfolding the data resulting in interpretation of much shorter ranges. Similar to nickel, the cobalt indicator semivariograms display a pattern of rotation anisotropy, increasing nugget effect, and decreasing indicator ranges with increasing indicator threshold. However, as a general comment the horizontal plane experimental variograms were poorly structured and the interpretations were based largely on the behaviour of the vertical, downhole results.

5 Estimation

The grade control sample locations were estimated from the exploration data with means of ordinary point kriging (OK) and multiple indicator kriging (IK) E-type estimates using the indicator thresholds discussed above (Journel, A.G., Huijbregts, C.J. 1978). Table 2 compares the grade control data statistics to the estimate made at each grade control location using the combinations of estimation method, data configuration and metal.

Stat	Nickel grade (%)					Cobalt grade (%)				
	Grade control	OK		IK		Grade control	OK		IK	
		In situ	Unfold	In situ	Unfold		In situ	Unfold	In situ	Unfold
Min.	0.07	0.16	0.15	0.32	0.32	0.001	0.001	0.001	0.001	0.001
Max.	2.67	2.06	2.00	2.07	2.12	0.887	0.887	0.887	0.887	0.887
Mean	0.85	0.85	0.85	0.87	0.87	0.058	0.055	0.060	0.056	0.061
Med.	0.81	0.83	0.83	0.84	0.85	0.040	0.052	0.055	0.052	0.056
S.D.	0.38	0.23	0.26	0.23	0.28	0.054	0.025	0.027	0.026	0.031
C.V.	0.44	0.26	0.31	0.26	0.32	0.944	0.452	0.459	0.464	0.503

Table 2. Summary statistics of exploration grade estimates compared to grade control

In terms of mine planning and the need to repay start-up capital expenditure in the early years of mining and processing, the corrected estimation of the amount of high-grade material at the exploration stage is critical to project feasibility. Despite the fact that the estimation results and input data are point values, pseudo grade volume curves have been generated for each estimate by assuming that each node represents an ore parcel of dimension 12.5 m E by 12.5 m N by 2 m RL. These curves give an insight into accuracy of high-grade volume estimates that can be expected from each estimation method and are shown in Figure 4.

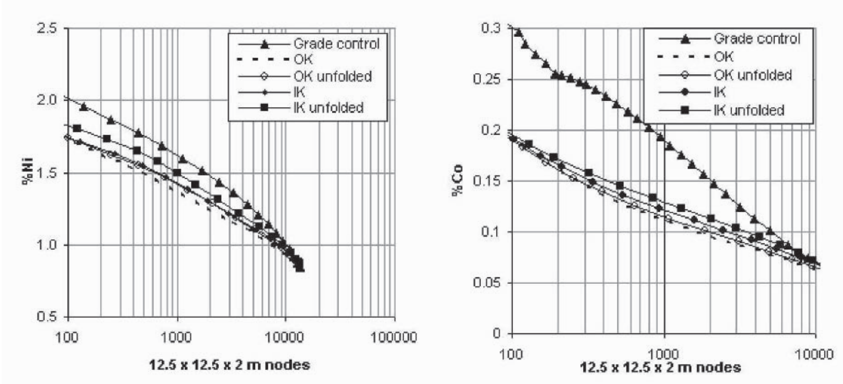


Figure 4. Pseudo grade tonnage curves for nickel (left) and cobalt (right)

6 Conclusions

The statistics in Table 2 reveal that for nickel, OK gives the most accurate estimate of the mean but the results plotted in Figure 4 show that the unfolded IK method is more accurate in estimation of the high grade ore. For cobalt, the best accuracy of both mean and high grade is also achieved for the combination of unfolding and IK estimation method albeit all methods poorly predict the amount of high grade cobalt.

References

Brand, N.W., Butt, C.R.M., Elias, M. (1998): Classification and features of Nickel Laterites. Cooperative Research Centre for Landscape Evolution and Mineral Exploration Report 74, CSIRO, Perth.

Dowd, P. A., Johnstone, S.A.W., Bower, J. (1988): The application of structurally controlled geostatistics to the Hilton Orebodies, Mt Isa, Australia. In: 21st Application of Computers and Operations Research In the Mineral (APCOM) Industry. A. Weiss (Ed), Society of Mining Engineers. (Vol 2), p 275-285.

Fazakerley, V.W., Monti R. (1998): Murrin Murrin nickel-cobalt deposits. In: Geology of Australian and Papua New Guinean Mineral Deposits, D. A. Beriman and D. H. Mackenzie (Eds), The Australasian Institute of Mining and Metallurgy. Melbourne. p 329-334.

Golightly, J. P. (1981). Nickeliferous Laterite Deposits. Econ. Geol., 75th Anniversary Volume, p 710-713.

Journel, A.G., Huijbregts, C.J. (1978). Mining Geostatistics. Academic. London.

Lambert, S. (2000): Geostatistical estimation and generalised spatial coordinate transformations. In: Geostats 2000, WJ Kleingeld and DG Krige (Eds), p 864-883, Cape Town.

McArthur, G.J. (1998): Using geology to Control geostatistics in the Hellyer deposit. Mathematical Geology, 20, 5. p. 342-366, Dordrecht.

Murphy, M. (1998): Murrin Murrin East resource estimation, Geostatistical Association of Australia Newsletter Sept 1998. [On-line] WWW: http://www.confed.com.au/gaa/.

Murphy, M, Bloom L and Mueller U (2002): Geostatistical optimisation of mineral resource sampling cost for a Western Australian nickel deposit. In Bayer U, et.al. (eds) IAMG 2002 Proceedings of the 8th Annual Conference of the International Association of Mathematical Geology, Terra Nostra Berlin, p.209-214.

Sahin, A Ghori, S.G., Ali, A.Z., El-Sahn, H.F., Hassan, H.M. and Al-Sanounah, A. (1998): Geological controls of variograms in a complex carbonate reservoir, Eastern Province, Saudi Arabia. Mathematical Geology, 30, 3. p 309-322, Dordrecht.

Sides, E.J., da Silva, F.J. (1996): Application of variable search prism orientation for improved geological control on grade estimation at the Neves-Corvo copper-tin mine, Proceedings of the Conference on Mining Geostatistics, Geostatistical Association of South Africa, 182-199, South Africa.

Wellmer F.W., Giroux, G.H. (1980): Statistical and geostatistical methods applied to the exploration work of the Nanisivk Zn-Pb Mine, Baffin Island, Canada. Mathematical Geology, 12, 4, p 321-337, Dordrecht.

MEASURES OF UNCERTAINTY FOR RESOURCE CLASSIFICATION

LUIS EDUARDO DE SOUZA, JOÃO FELIPE C.L. COSTA and JAIR C. KOPPE
Mining Engineering Department, Federal University of Rio Grande do Sul
Av. Osvaldo Aranha, 99/504, 90035-190, Porto Alegre/RS, Brazil

Abstract. For many decades the mining industry regarded resource estimation and classification as a mere calculation requiring basic mathematical and geological knowledge. Often uncertainty associated with tonnages and grades were either ignored or mishandled. With initiatives to establish international standards for classifying mineral resources and reserves, it is important to establish the level of confidence in the results and correctly assess the error. Among geostatistical methods, Ordinary Kriging (OK) is probably the one most used for mineral resource estimation. It is known that OK variance is unable to recognize local data variability, which is an important issue when heterogeneous mineral deposits with higher and poorer grade zones are being evaluated. This study investigates alternatives for computing estimation variance from ordinary kriging weights that account for both the data configuration and the data values. These estimation variances are then used to classify resources based on confidence levels and their results are compared with those obtained by OK variance. The methods are illustrated using an exploration drill hole data set from a large Brazilian coal deposit. The results show the differences in tonnages within each class of resources when different measures of uncertainty are used.

1 Introduction

The mining industry has already recognized and established standards for resource evaluation and classification but now, with the increasing internationalization of mining companies, the development of internationally acceptable standards for this classification has become relevant.

Since 1994 the Council of Institutions of Mining and Metallurgy (CMMI), an international entity that congregates institutions from the United States (SME), Australia (AusIMM), Canada (CIM), United Kingdom (IMM) and South Africa (SAIMM), has proposed a set of definitions for the reporting and classification of mineral resources and reserves. These definitions were adopted later by a committee established in 1998 by the United Nations thus granting it truly international recognition.

The main mineral resource classification systems adopted worldwide are essentially based on sampling spacing, geological confidence and economical viability. These

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 529-536.

systems define classes of resources based on a degree of certainty associated with estimated tonnages and grades. Classes of in situ coal (measured, indicated and inferred) are defined based on the spatial distribution of the samples and the uncertainty associated with tonnages calculated for a given deposit or part of it. Thus, classifying in situ coal or coal resources requires the definition of the uncertainty associated with the estimate. However, what is not stated in the classification systems is how uncertainty should be measured, and even the JORC rules, such as in Table 1, provides a minimum necessary data density, but it does not specify or advise any estimation algorithm or how uncertainty should be assessed.

Classes of resources	Maximum extrapolation distance	Maximum spacing between points of observation[1]	Degree of uncertainty
Measured	500 m	+ 1 km; < 500 m	0 - 10%
Indicated	1,000 m	+ 2 km; < 1 km	10 - 20%
Inferred	2,000 m	+ 4 km	> 20%

Table 1. Classes of resources based on sampling spacing defined by the JORC system.
[1] The first distance is the acceptable limit and the second is the normally used distance.

Since classification codes are not prescriptive regarding the estimation method used, several geostatistical approaches have been suggested, mainly because these techniques provide a short, unambiguous identification of resources/reserves categories. Geostatistical estimate methods are suggested in most codes and these methods have become the accepted standard models for mineral resource estimates. Several geostatistical methods can be used to estimate and assess uncertainty. Among them ordinary kriging is probably the most widely used mainly due its specific features related simplicity and reliable estimates (Matheron, 1963; David, 1977; Journel, 1983; Isaaks and Srivastava, 1989). However, the geostatistical literature has discussed the misuse of ordinary kriging variance as an accurate measure of uncertainty, mainly because it is only variogram dependent and not data-value dependent, taking into account only the spatial arrangement of the samples, and consequently ignoring the local variability (Journel, 1986).

This study investigates some of the proposed alternatives that have been proposed to the kriging variance. Two distinct approaches to calculate estimate variance via ordinary kriging weights were used: (i) the interpolation variance (Froidevaux, 1993; Yamamoto, 1999) and (ii) the combined variance (Arik, 1999). The obtained results are compared with those derived from OK variance. All the methodologies were repeated to four different block sizes, trying to identify its influence on the estimates, as it is known the larger the block size the smaller the associated variance (Krige, 1996).

The estimate and the subsequent classification of resources into different classes or categories, according to the possible variations of these resources must provide a model that quantifies the risk on each category. A comparative study was carried on using an exploration data set from a large Brazilian coal deposit and the results show the impact in both tonnages and error in each resource category when the alternatives to uncertainty assessment are used.

2 Case study

The deposit object of this study is located in the southern Santa Catarina coal basin and it has been exploited since the early 1900´s. The depositional environment imposed a particular geometry to this deposit with a longer continuity for coal thickness along the major axis of deposition and a short range along the orthogonal direction.

Since the samples used for resource assessment should be representative and present a high degree of confidence, all drill holes with poor reliability in terms of core yield or logging criteria were omitted from the deposit modeling. Thus, from the original 471 ddh, 340 were kept for thickness estimate purpose and 236 for the specific gravity. As the collars were not regularly spaced, a declustering procedure (Deutsch and Journel, 1998) was used to obtain a representative statistic for the entire area. In the sequence, spatial continuity analysis was carried out modeling the major and minor directions of anisotropy. A two-structure spherical variogram (Sph) model [γ(h)] was estimated from the experimental variogram points for the two variables as:

$$\gamma(h) = 0.045 + \left[0.057 Sph_{(1)} \left[\frac{hN-S}{467m}, \frac{hE-W}{233m} \right] + 0.233 Sph_{(2)} \left[\frac{hN-S}{6131m}, \frac{hE-W}{1897m} \right] \right] \qquad (1)$$

for coal thickness, and:

$$\gamma(h) = 0.0015 + \left[0.0033 Sph_{(1)} \left[\frac{hN135E}{676m}, \frac{hN45E}{478m} \right] + 0.0029 Sph_{(2)} \left[\frac{hN135E}{3190m}, \frac{hN45E}{2440m} \right] \right] \qquad (2)$$

for specific gravity, where $Sph = \begin{cases} \frac{3}{2}\frac{h}{a} - \frac{1}{2}\left(\frac{h}{a}\right)^3 & if\ h \le a \\ sill & if\ h > a \end{cases}$, and a is the range of the variogram.

Coal thickness anisotropy coincides with the main axis of the coal basin. Specific gravity anisotropy directions are oriented according to the shoreline which used to divide the lacustrine/marine environment where this coal deposit was formed. Pyrite concretions presence and consequently the increase in the specific gravity is controlled by this shoreline orientation. Therefore, these two geological attributes not necessarily should have their major axis of anisotropy coincident.

The main parameters used for modelling the deposit into mineable blocks using ordinary kriging were: (i) minimum of 4 and maximum of 24 data located in the local neighbourhood of a given block being interpolated, (ii) 64 points used to discretise the block and obtain an average estimate of it, (iii) searching for samples around the block within the variogram ranges defining an ellipsoidal search, (iv) four different block sizes (175 x 175 m, 350 x 350 m, 525 x 525 m, and 700 x 700 m), (v) searching for samples in the local neighbourhood of a block dividing the search ellipsoid into octants. The variograms and the parameters used for kriging were cross validated (Isaaks and Srivastava, 1989).

Based on the available data and the JORC code standards for extrapolation distance and distance between samples (Table 1), the boundaries defined by geometric definitions were established. Adopting the usually recommended values, the areas covered by a single sample were disregarded for framing in measured or indicated coal in situ categories.

3 Assessing uncertainty

3.1 KRIGING VARIANCE

Ordinary kriging produces a set of estimates for which the variance of the errors is minimized through the use of the Lagrange multipliers and is usually referred to as the ordinary kriging variance:

$$\sigma^2_{OK}(u) = C(0) - \sum_{\alpha=1}^{n(u)} \lambda^{OK}_{\alpha}(u)\, C(u_{\alpha} - u) - \mu_{OK}(u) \tag{3}$$

where C(0) is the a priori variance of the data, λ^{OK}_{α} is the weight calculated for each datum in the neighbourhood of u, C(u_α - u) is the covariance from each datum and the location u and μ_{OK} is the Lagrange multiplier (Matheron, 1963). The kriging variance computed for a given point or block being estimated is essentially independent of the data values used in the estimation and it does not measure uncertainty, but just the spatial configuration of local data used to make the estimate. The link between kriging variance and data values is just through the variogram, which is global rather than local in its definition (Arik, 1999; Journel, 1986; Isaaks and Srivastava, 1989; Yamamoto, 1999).

3.2 INTERPOLATION VARIANCE

Yamamoto (1999) proposes the interpolation variance as the weighted average of the squared differences between data values and the OK estimate according the following expression:

$$s_0^2 = \sum_{i=1}^{n} \lambda_i \left[z(x_i) - z^*(x_0)\right]^2 \tag{4}$$

where λ_i are the ordinary kriging weights, *z(x$_i$)* are the n neighbor data close to the unsampled location *(x$_0$)*, and $z^*(x_0)$ is the block estimate. This expression is exactly the same that proposed by Froidevaux (1993). It is data-value dependent and this definition requires all weights be positive since any negative weight could lead to a negative interpolation variance. There are several available solutions for avoiding negative weights and they can be basically divided into two types: (i) ordinary kriging weights can be constrained to be positive before solution of the ordinary kriging system (Barnes and Johnson, 1984; Herzfeld, 1989), or (ii) correct the negatives after kriging (Froidevaux, 1993; Journel and Rao, 1996; Deutsch, 1996). This study adopted the procedure proposed by Deutsch (1996), and his solution was implemented in the kriging routine *kt3d* of GSLIB (Deutsch and Journel, 1992).

3.3 COMBINED VARIANCE

Arik (1999) suggests an alternative measure to assess the uncertainty that is basically a combination of the kriging variance and the variance of the weighted average block value based on the data values used. The second is defined as:

$$\sigma_W^2 = \sum_{i=1}^{n} w_i^2 \left[Z_0 - z_i\right]^2 \qquad i = 1,\ldots,n \ \ (n > 1) \tag{5}$$

where n is the number of data used, w_i are the ordinary kriging weights corresponding to each datum, Z_0 is the block estimate, and z_i are the data values. If there is only one datum, σ_w^2 is set to σ_{OK}^2 . This component, called by Arik (1999) the local variance of the weighted average, is then used to calculate the combined variance (σ_{CV}^2) as follows:

$$\sigma_{CV}^2 = \sqrt{(\sigma_{OK}^2 \times \sigma_W^2)} \tag{6}$$

Suppose the example in Figure 1. For sake of simplicity, there are only seven data points surrounding the point 0 to be interpolated. Using an exponential variogram model with an isotropic range of 10, sill of 10, and nugget of 0, ordinary kriging was used to calculate the value at location 0, resulting in 592.7 with a kriging variance of 8.96.

If one changes the data set according to Figure 1b, keeping everything else the same, the new estimate would be 550.0. The kriging variance remains the same at 8.96, since the variogram parameters and data configuration were the same as the first run. Table 2 summarizes the results. One can observe the alternative variances reflect local variability.

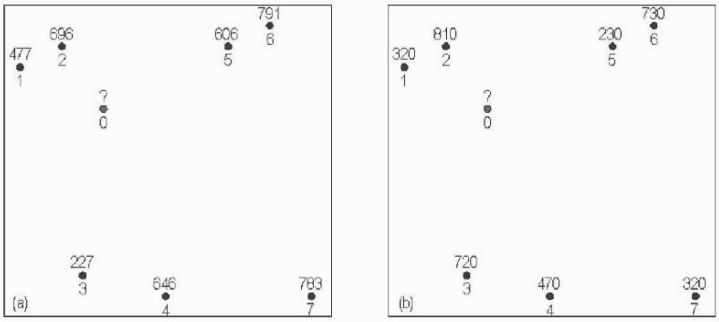


Figure 1. Sample data and location 0 (a) extracted from Isaaks and Srivastava (1989). The same data configuration with a different set of values (b).

Samples	Kriging Variance	Weighted Average	Combined Variance	Interpolation Variance
(a)	8.96	4114.6729	191.9667	28551.7148
(b)	8.96	11769.0791	324.6606	56772.0156

Table 2. Variances for the two data sets presented in Figure 1.

3.4 ERROR DEFINITION

Block models of four different sizes were defined for both variables using ordinary kriging. These models were validated and the results, including the estimated variance, used to define confidence intervals. For each block, the coal accumulation (t/m^2), expressed as a product of the thickness by density has its variance evaluated and expressed using (David, 1977):

$$\frac{\sigma_{xy}^2}{(xy)^2} = \frac{\sigma_x^2}{x^2} + \frac{\sigma_y^2}{y^2} + 2\rho_{xy}\frac{\sigma_x}{x}\frac{\sigma_y}{y} \tag{7}$$

where σ_{xy}^2 is the variance of the product, ρ_{xy} is the coefficient of correlation, σ_x and σ_y are the standard deviations, x and y are the estimated block values to thickness and specific gravity. The third term in Equation (7) is null since the correlation between density and thickness is insignificant (Figure 2).

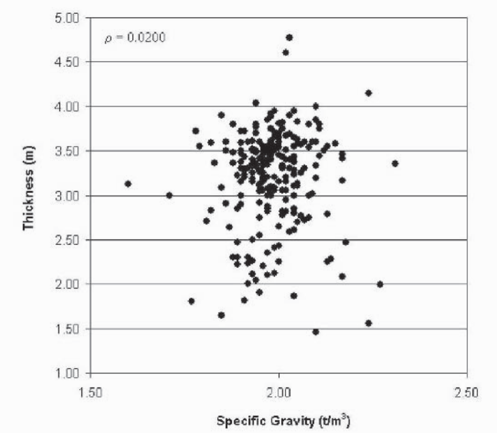


Figure 2. Scatterplot for specific gravity (t/m^3) versus thickness (m). Note the absence of correlation.

Assuming a Gaussian distribution to the error, the confidence interval with 95% probability of containing the mean can be approximated using (David, 1977):

$$\overline{t/m^2}(n) \pm t_{n-1,1-\alpha/2}\sqrt{\frac{\sigma_{xy}^2}{n}} \tag{8}$$

where $\overline{t/m^2}$ is the inferred mean to n data, $t_{n-1,1-\alpha/2}$ is the $1-\alpha/2$ superior critical point for the t distribution with n - 1 degrees of freedom.

Thus, the global error for each coal in situ category was obtained using each accumulation block value as a weight to the block error, according to the theory of errors presented by Caputo (1969).

4 Discussion and conclusions

The approach presented was repeated for each one of the alternative variance discussed and for each block size tested. The results on either tonnages or error for each category are showed in Tables 3 and 4.

Block size	Tonnages of coal in situ (t x 10^6)		
	Measured	Indicated	Inferred
175 x 175 m	237.62	126.62	188.11
350 x 350 m	239.54	124.70	189.86
525 x 525 m	241.51	132.54	187.47
700 x 700 m	234.74	116.07	197.08

Table 3. Calculated tonnages of coal in situ for different block sizes.

Block size	Variance	Error (%)		
		Measured	Indicated	Inferred
175 x 175 m	σ_{OK}^2	4.43	7.61	11.33
	s_0^2	7.54	13.37	17.77
	σ_{CV}^2	3.12	5.61	8.34
350 x 350 m	σ_{OK}^2	3.64	6.92	10.53
	s_0^2	7.58	13.48	17.61
	σ_{CV}^2	2.81	5.38	8.01
525 x 525 m	σ_{OK}^2	3.19	6.33	10.46
	s_0^2	8.01	13.49	17.79
	σ_{CV}^2	2.68	5.14	7.96
700 x 700 m	σ_{OK}^2	2.76	5.84	9.47
	s_0^2	7.98	12.30	17.28
	σ_{CV}^2	2.44	4.58	7.49

Table 4. Confidence limits for the coal in situ calculated tonnages obtained via kriging variance (σ_{OK}^2), interpolation variance (s_0^2), and combined variance (σ_{CV}^2).

In Table 3, it is observed that the estimated tonnages have changed for the different block sizes tested, these variations were generally small, and only in the category of indicated in situ coal was the variation about 12%, with the increase of the block size. This seems to be related with a more complex geometry for this class as well as a different adherence that each size has regarding the geometric boundaries that define the resources categories. Table 4 shows that the calculated values of error using the methodology proposed by Yamamoto (1999) are substantially higher than the calculated ones with kriging variance as well as combined variance. Several blocks were classified as measured resources according to the geometric criteria, but could not be classified as indicated or even inferred due to uncertainty criteria. In this study, these blocks were not removed from the resources inventory or re-arranged into different categories, however if this variance was used for classifying a reduction on the resources would occur.

The fact that the case study consists of a tabular orebody, extremely continuous spatially and with abundance of information may have contributed to attenuate the differences between the results obtained via kriging variance and combined variance, and these factors may explain the small differences in terms of tonnages and error with the increment of the block size. Even so, there are significant differences in the calculated error using each one of the alternative variances.

Ordinary kriging (OK) variance (or its square root, the standard error) has been largely used as a measure for spread of the estimates, but since this parameter depends only on (i) the spatial continuity of the data and (ii) the spatial configuration of the observations, the error calculated using OK variance will be independent from the data values imposing severe limitation on its use. Therefore, the use of alternative measures of uncertainty allow a more accurate and coherent response. These measures for the uncertainty eliminate the subjectivity of using a fixed or empirical range of influence as discriminating factor among the categories of resources that do not respect the singularity of each mineral deposit.

References

Arik, A. An Alternative Approach to Resource Classification, *in Proceedings of the 28th International Symposium on Computer Applications in the Mineral Industries (APCOM'99)*, Colorado School of Mines, Golden, Colorado USA, 1999, p. 45-53.

Barnes, R.J. and Johnson, T.B. Positive Kriging, *Geostatistics for Natural Resources Characterization*, Reidel, Dordrecht, 1984, p. 231-240.

Caputo, H.P. *Matemática para a Engenharia*, Ed. Ao Livro Técnico S.A., Rio de Janeiro, 1969, 416 p.

David, M. *Geostatistical Ore Reserve Estimation, Developments in Geomathematics 2*, Elsevier Scientific Publishing Company, Amsterdam, 1977, 364 p.

Deutsch C.V. and Journel, A.G. *GSLIB: Geostatistical Software Library and User's Guide*, Oxford University Press, New York, USA, 1998, 335 p.

Deutsch, C.V. Correcting for Negative Weights in Ordinary Kriging, *Computers & Geociences*, vol. 22, no. 7, 1996, p. 765-773.

Froidevaux, R. Constrained Kriging as an Estimator of Local Distribution Functions, in Capasso, V., Girone, G., and Posa, D., eds., *in Proceedings of the International Workshop on Statistics of Spatial Processes: Theory and Applications*, Bari, Italy, 1993, p. 106-118.

Herzfeld, U.C. A Note on Programs Performing Kriging with Nonnegative Weights, *Mathematical Geology*, vol. 21, no. 3, 1989, p. 391-393.

Isaaks, E.H. and Srivastava, M.R. *An Introduction to Applied Geostatistics*, Oxford University Press, New York, USA, 1989, 561 p.

Journel, A.G. Non-Parametric Estimation of Spatial Distributions, *Mathematical Geology*, vol. 15, no. 3, 1983, p. 445-468.

Journel, A.G. Geostatistics: Models and Tools for the Earth Sciences, *Mathematical Geology*, vol. 18, no. 1, 1986, p. 119-140.

Journel, A.G. and Rao, S.E. Deriving Conditional Distributions from Ordinary Kriging, *Stanford Center for Reservoir Forecasting (Report No. 9)*, Stanford University, 1996, 25 p.

Krige, D.G. A Practical Analysis of the Effects of Spatial Structure and of the Data Available and Accessed, on Conditional Biases in Ordinary Kriging, *in Proceedings of Geostatistics Wollongong´96*, vol. 2, 1996, p. 799-810.

Matheron, G. Principles of Geostatistics, *Economic Geology*, no. 58, 1963, p. 1246-1266.

Yamamoto, J.K. Quantification of Uncertainty in Ore-Reserve Estimation: Applications to Chapada Copper Deposit, State of Goiás, Brazil, *Natural Resources Research*, vol. 8, no. 2, 1999, p. 153-163.

INCORPORATING UNCERTAINTY IN COAL SEAM DEPTH DETERMINATION VIA SEISMIC REFLECTION AND GEOSTATISTICS

VANESSA C. KOPPE, FERNANDO GAMBIN, JOÃO FELIPE C. L. COSTA, JAIR C. KOPPE, GARY FALLON and NICK DAVIES
Department of Mining Engineering, Federal University of Rio Grande do Sul
Brazil, RS, POA, Oswaldo Aranha Avenue, 99- 506

Abstract. Modelling mineral deposits requires the use of all possible source of information. Traditionally, core samples from borehole are the most used way to access the ore body, however this method is expensive and provides information restricted to a close neighbourhood within the sample location. Continuity between sampled points needs to be inferred in order to infill values among bore hole locations using interpolation techniques. In contrast, geophysical methods including seismic reflection provide data at much closer intervals, thus approximating continuous sampling along a seismic section. These data are then used to infer spatial continuity, for example the fault of a coal seam in between bore holes. Wave travel time along the seams is recorded by seismic survey at a dense grid. Additionally, sonic wave velocity logged along boreholes can be interpolated at a dense grid. Sonic Logging provides direct and continuous measurements of the sonic wave velocity at all seams down the holes logged. Therefore, this logged sonic velocity can be simulated within a dense grid compatible to the time grid. Multiple velocity grids (equally probable models) are generated within the simulation framework. In combining both grids, i.e. velocity and time, seam depth can be obtained. Consequently risk in depth determination for each seam due to velocity uncertainty can be assessed. Both data types (time and sonic) are subject to various sources of error. Currently, velocity is indirectly determined using processed seismic data, which may breed errors in geologic sections interpretation. The present paper will show results from a Sonic Logging velocity simulation and its uncertainty determination, in order to use the results in calculating seam depth via seismic reflection and additionally provided a measure for error in this parameter. A case study in a major coal deposit illustrates the procedure.

1 Introduction

Modelling mineral deposits is based on a conceptual geological model and on readings derived from samples sparsely taken within this deposit. Usually, the samples are collected by diamond drill holes (core sampling). However, this sampling technique is very expensive (~ US$ 100/m) and provides restricted amount of information, imposing all sort of difficulties in reducing the uncertainties associated with the estimation of geological attributes within the mineral deposit.

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 537-542.

Geophysical methods like seismic are a direct method of more coarsely, but cheaply sampling, which can be very accurate in favourable conditions. Seismic methods measure the mechanical wave propagation time from a source to a receiver within rocks. This time can be related to the geometry of seam layers or bed rocks. The depth of the beds is obtained multiplying the time the wave took to travel by the wave's velocity along that rock. Accurate velocities can be obtained from sonic logging sampling. This method can collect velocities at various points along borehole walls.

This study aims to estimate seismic wave velocity in a 3D grid, based on sonic logging data (note that seismic waves velocities can be correlated to sonic logging data). Sonic logging measures slowness, which is an attribute that is defined as the inverse of compressional velocity. For sake of this study slowness will be called and treated as velocity samples. The estimates will be generated using geostatistical methods. The development of an appropriate modelling methodology could facilitate a better depth conversion from seismic data. If the sonic velocity data were converted to a rock mass parameter then a more accurate geotechnical model will also be constructed.

Geostatistics comprises a collection of tools used to estimate values of any attribute of a mineral deposit at unknown locations, supported by its spatial continuity model. The geostatistical tools used on this study include ordinary kriging (Matheron, 1963) and sequential Gaussian simulation (Isaaks, 1990). Sequential Gaussian simulation provides a method of assessing the uncertainty associated with the estimated velocity, which can also be approximate via ordinary kriging variance. However the later must be used with caution given certain limitations (Goovaerts, 1997).

The methodology presented is illustrated by a case study where the uncertainty related to the velocity is determined as well as the corresponding uncertainty of the coal seam depth. Sampling errors generated in seismic and sonic logging also contribute to depth uncertainty, however these errors were not considered in this study. The target coal seam is approximately 210m below the surface. The seam has an average thickness obtained from the 60 core samples) of 2.1m and is extracted by longwall retreat. The overlying stratigraphy is a siliclastic sequence containing at least 9 thinner but of variable thickness coal seams.

2 Case Study

2.1 DATA SET

A coal deposit was used and velocity samples were collected from 60 logged boreholes. Each borehole logged was sampled at 5 cm intervals along 300 m (average hole length). The dataset comprises 228851 sonic wave velocity samples (unit µs/ft). These samples were obtained by geophysical logging along 60 core and non-core drill holes (Figure 1).

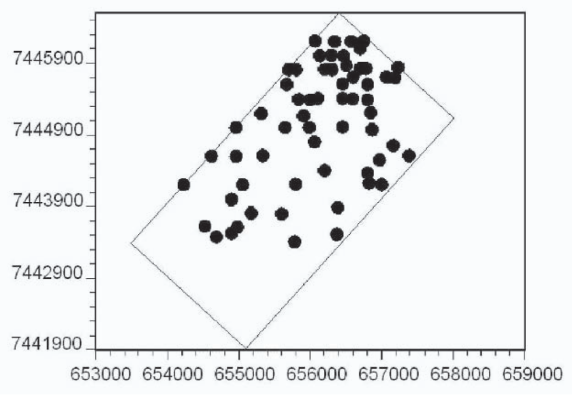


Figure 1- Location map for 60 borehole collars and 3D seismic survey boundary (black line).

2.2 VARIOGRAPHY

The vertical experimental variogram for sonic wave velocity samples showed a fast increase at the first meters due to short scale high variability of this attribute. The horizontal omnidirectional experimental variograms for sonic wave velocity samples showed a high degree of continuity as expected since all the readings tend to belong to the same strata along the horizontal plan.

2.3 KRIGING

The sonic wave velocity attribute was interpolated using ordinary kriging (Matheron, 1963). Figure 2 shows vertical sections sliced from the kriged block model. The smoothing effect is evident in the kriged 3D block model. In a shallow dipping sedimentary environment one might expect near horizontal layering to be evident. While stratigraphic units are correlated with horizontal distance there are changes in their physical property distribution.

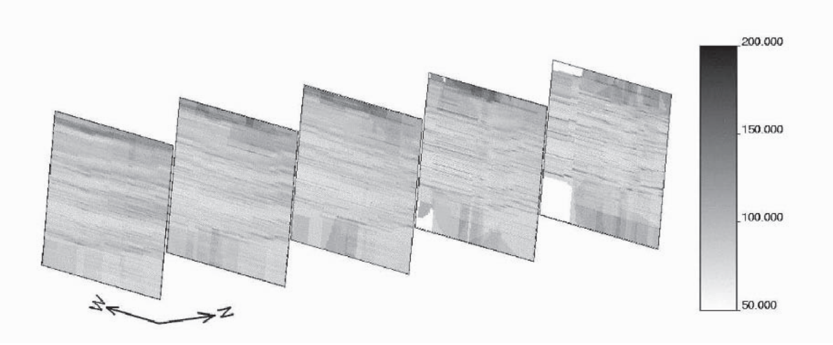


Figure 2- Vertical sections (longitudinal views along XZ plan) at various North (Y) coordinates extracted from the kriged block model. Gray scale represents kriged sonic velocity (μs/ft).

Figure 3 shows sections along XZ plan sliced from the 3D block model representing the kriging standard deviation of each block. Kriging standard deviation calculated at blocks near drill holes provide low values as is observed in Figure 3. Light grey vertical lines (lowest standard deviation values) identify borehole locations.

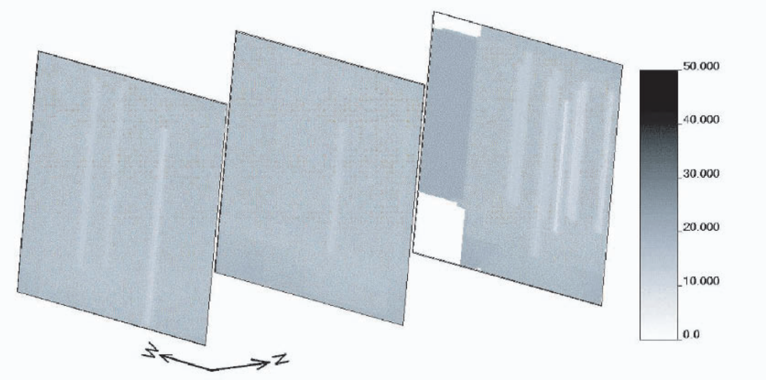


Figure 3- Vertical sections (longitudinal views along XZ plan) at various North (Y) coordinates plotting the standard deviation at every block resulting from kriging. Gray scale represents sonic velocity kriging standard deviation (μs/ft) at each block.

2.4 SIMULATION

Sequential Gaussian simulation (SGS) (Isaaks, 1990) was selected to be used in this case study. SGS provides realizations (maps) (Deutsch and Journel, 1998) of sonic velocity, where each realization is a possible representation for the attribute being studied. Simulations were conditioned to the 222648 samples collected along the 60 logged boreholes. The number of realisations (20) was considered enough for uncertainty assessing as at this number the ergodic fluctuations on the global mean reached a steady state, i.e. the variance of the mean stabilized. Figure 4 illustrates the same sections as in Figure 2 and shows the vertical sections sliced along XZ plans extracted from a 3D block model obtained by simulation. A granular texture on the grey scale maps is clearly noticed.

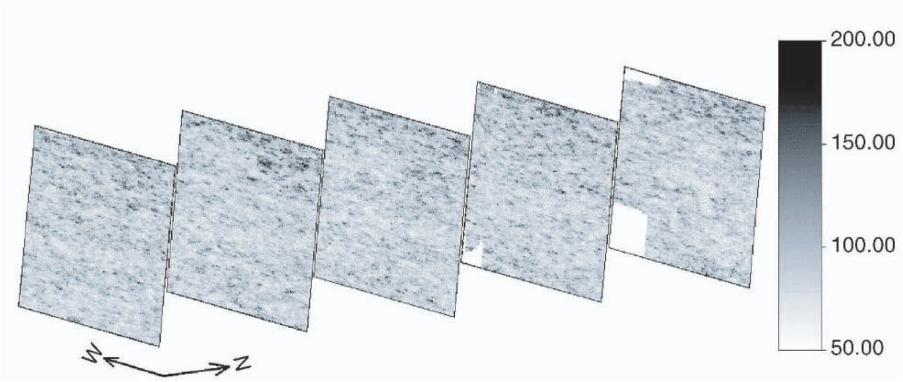


Figure 4- Vertical sections (longitudinal view XZ plan) at various North (Y) coordinates for the sonic velocity calculated at each block of one simulation. Gray scale represents simulated sonic velocity (μs/ft).

Conditional standard deviation was calculated at each simulated node using the values obtained at this node resulting from different realizations.

2.5 ERROR DETERMINATION

The uncertainty about kriged and simulated values can be quantified as the error. Assuming the error follows a normal distribution, the error interval can be calculated using kriging standard deviation or conditional standard deviation for the simulated values as follows (Christman, 1978):

$$\text{Error} = \pm\ t_{\frac{\alpha}{2},n-1} \frac{\sigma}{\sqrt{n}}$$

where σ is the standard deviation of the values; n is the number of values; $t_{\alpha/2,\ n-1}$ is parameter obtained from t-student distribution which depends on the confidence interval (1-α) and on degrees of freedom (n-1).

The error at each location derived from the kriged or simulated models were determined using confidence interval with 95% probability. It means there is 95% probability for the estimated value to be included in the confidence interval. Figure 5 shows the histograms for the error at all nodes generated by kriging and simulation. Visually the errors obtained via kriging have less variance (smoother) than the errors calculated via simulation. Statistically the error mean and median for the simulation method are lower. Practically, for a 2.1m thick seam at approximately 210m depth this difference is greater than the seam thickness. This difference is due to the smooth effect associated with interpolation methods. For this reason, simulation measure of uncertainty is larger than the one provided via kriging. The ability to improve the depth estimation by greater than a seam thickness is considerable for providing an accurate geological model to mine from.

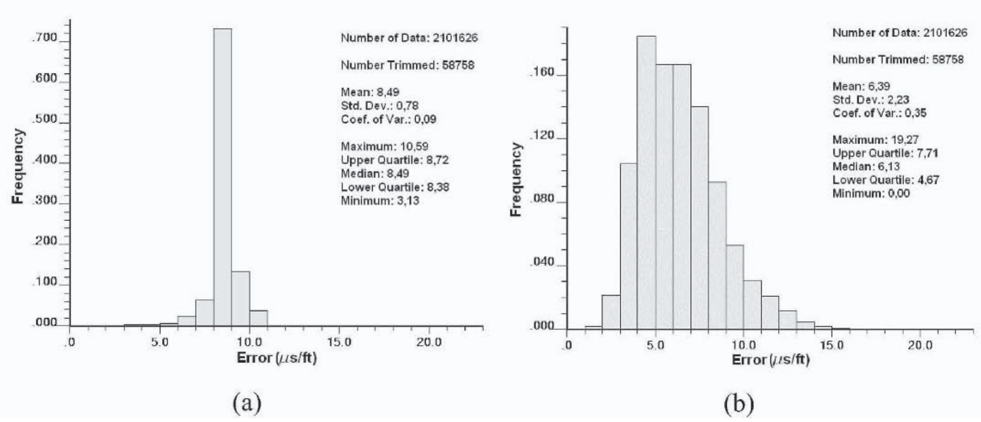


Figure 5- (a) Histogram of error measure at kriged nodes. (b) Histogram of error measure at all simulated nodes.

3 Conclusions

This study was aimed at estimating seismic wave velocity and its associated uncertainty at every node of a 3D grid, based on sonic logging data. The 3D model for velocity was estimated using ordinary kriging and sequential Gaussian simulation.

Ordinary kriging produced the best estimate at a price of smoothing the interpolated values and consequently the error forecasted. Sequential Gaussian simulation (20 realizations) produced 20 estimates at each grid node (one estimate for each realization). These simulated models globally resemble the real mineral deposit as they reproduce ergodically its spatial continuity.

Assuming a normal distribution for the error, a value at each grid node resulting from interpolating using ordinary kriging and sequential Gaussian simulation was calculated. Simulation produced a larger error interval (due to a larger space of uncertainty) but an overall lower mean of errors hence SGS is recommended as a process to derive a sonic velocity model.

References

Christman, R.U., *Estatística Aplicada*, Edgard Blücher Ltda, 1978.

Deutsch, C.V. and Journel, A.G., *GSLIB: Geostatistical Software Library and User's Guide*, 2nd Edition, Oxford University Press, 1998.

Goovaerts, P., *Geostatistics for Natural Resources Evaluation*, Oxford University Press, 1997.

Isaaks, E.H., *The Application of Monte Carlo Methods to The Analysis of Spatially Correlated Data*, Ph.D. Thesis, Leland Stanford Junior University, 1990.

Matheron, G., Principles of Geostatistics, *Economic Geology*, no. 58, 1963, p. 1246-1266.

IMPLEMENTATION ASPECTS OF SEQUENTIAL SIMULATION

STEFAN ZANON and OY LEUANGTHONG
Department of Civil & Environmental Engineering, 220 CEB,
University of Alberta, Canada, T6G 2G7

Abstract. Sequential simulation is used throughout the natural resources industry to construct multiple equiprobable numerical models. The sequential methodology is straightforward, but some implementation details require further explanation. This paper explores some of the implementation issues associated with choice of: simulation path, search strategies, number of conditioning data, and the affect of ergodic fluctuations under the Gaussian assumption.

1 Introduction

Sequential simulation (SS) (Johnson, 1987; Journel, 1993) is a stochastic modelling approach that yields multiple realizations based on the same input data. This data could be either continuous or categorical. Depending on the data type, sequential indicator simulation (SIS) (Gomez-Hernandez and Srivastava, 1993), sequential Gaussian simulation (SGS) (Isaaks, 1990), or direct sequential simulation (DSS) (Xu and Journel, 1994; Caers, 2000; Soares, 2001) will be used. This suite of simulation techniques has greatly expanded the tools that are available for building stochastic models, while injecting more variability than their kriging counterparts. The SS work-flow can be described in four basic steps:

1. Choose the stationary domain.
2. Define a path to visit every location.
3. At each location:
   a) search to find nearby data and previously simulated values,
   b) calculate the conditional distribution, and
   c) perform Monte Carlo simulation (MCS) to obtain a single value from the distribution.
4. Repeat step 3 until every location has been visited.

For SIS and SGS, a pre- and post-processing data transformation step is required. In the SIS case, data are transformed into indicator variables; in SGS, data are transformed to be Gaussian via a quantile transform or a Gaussian anamorphosis (Chilès and Delfiner, 1999). The above methodology produces one possible re-

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 543-548.

alization, and more realizations can be created by choosing a different random path.

The theory behind each form of SS has been explained numerous times (Isaaks, 1990; Journel, 1993; Goovaerts, 1997; Chilès and Delfiner, 1999; Deutsch, 2002), but it is the details of sequential simulation that warrant further explanation. In the publicly available GSLIB (Deutsch and Journel, 1998) programs like SGSIM or SISIM, for Gaussian and indicator simulation, respectively, the user must specify how key aspects of the simulation will be performed. These decisions can greatly affect the resulting model and the associated CPU requirements. A better understanding of these decisions will help the user to improve their models while balancing efficiency with accuracy.

2 Data and Transformation

Before simulation can be performed the model area must be defined and the input data identified. In general, the data must come from a single underlying distribution. The mean, variance, and higher order statistics are then assumed stationary throughout the area, that is, $E\{Z(\mathbf{u})\} = m$ and $Var\{Z(\mathbf{u})\} = E\{[Z(\mathbf{u}) - m]^2\}$ (Journel and Huijbregts, 1978). If stationarity is violated, the mean and variance will change with location. A trend model can be used to describe these regional changes, and either the trend is removed to create stationary residuals or the trend is used as secondary data in a specialized form of kriging (Deutsch, 2002).

The underlying distribution of the modelling area, as described by the cumulative distribution function (cdf), should be reproduced in every simulation. This cdf is typically determined from the input data, but the data collection process is rarely performed to fairly sample the underlying distribution. To correct this, declustering (Isaaks and Srivastava, 1989; Goovaerts, 1997) can help to remove the affects of non-representative sampling or a reference distribution, based on some secondary data or expert knowledge, can be used as a target distribution.

SGS requires the data to be standard Gaussian with zero mean and unit variance. To achieve this, the input data cdf is transformed through the quantiles to any other cdf. This one-to-one quantile transformation is reversible and allows the mean, variance and shape of the distribution to be changed while preserving the rank of the data (Journel and Huijbregts, 1978). Spikes in the cdf prevent the one-to-one quantile transform and despiking (Verly, 1984) will be required.

In SGS, the original distribution is reproduced by reversing the above transformation. This back transformation requires the data to follow the standard normal distribution; however, statistical fluctuations are inherent in simulation. Fluctuations in the mean and variance should be reasonable and unbiased. Small deviations in normal space can be magnified after back transformation, particularly if the original data follow a skewed distribution.

For example, consider a lognormal distribution, with mean and standard deviation of 8.0, and its corresponding normal score distribution after transformation, $N(0, 1)$. The effect of deviations from the standard normal distribution can be assessed by generating *near* standard normal distributions and reversing the above transformation. For this exercise four scenarios will be considered using standard deviations of 0.8 and 1.2, and means of -0.1 and 0.1. When both the mean and standard deviation are low in normal space, $N(-0.1, 0.8)$, the mean and standard deviation in original units are 6.64 and 5.96, respectively. Using the same mean with the higher standard deviation, $N(-0.1, 1.2)$, the original space mean and standard deviation are 8.48 and 9.37. The remaining two scenarios for the high mean, $N(0.1, 0.8)$, and $N(0.1, 1.2)$, result in an original space mean of 7.92 and 9.76, and standard deviations of 6.64 and 10.06, respectively. This example shows how sensitive the summary statistics in original space are to the ergodic fluctuations inherent in stochastic simulation in normal space for a skewed distribution. One proposed solution to mitigate these effects is to apply a standard transform to the simulated values (Journel and Xu, 1994).

3 Simulation Path

At every unsampled location, SS should use all available input data and previously simulated values as conditioning data. No assumption is made about the order in which these locations are visited, but the order will influence the final model. To minimize this influence, the starting location and path should be random (Isaaks, 1990; Tran, 1994). Over multiple realizations the structure in the model will be based on the data and not an artifact of the path.

Alternatives such as the regular path and spiral path (McLennan, 2002) have been considered, but any perceived benefits in CPU efficiency or input data propagation come at the expense of variogram reproduction and accuracy of the local distribution. These paths, along with the random path, can suffer from poor long range variogram reproduction, since the nearby data will preferentially be used as conditioning data. To avoid this problem, a multiple grid search (Tran, 1994) can be incorporated into the random path to improve variogram reproduction.

4 Searching for Local data

Before kriging can be implemented, a search is performed to identify surrounding conditioning data. The user limits this search by specifying a search radius in each principle direction, where the radii should equal or exceed the variogram range to ensure adequate variogram reproduction. Data beyond this range will provide limited information to the kriging estimate.

It is common practice to assign the input data to the grid nodes. This will *exactly* reproduce the input data in the final model and allow the covariance to be quickly calculated using a covariance look-up table. The disadvantage is that only a single data is retained in each grid cell and the remaining data are only used to establish the reference distribution. Also, input data cannot be preferentially used over previously simulated values. The spiral search (Deutsch and Journel, 1998) uses the covariance look-up table to develop a search path based on the decreasing correlation of the surrounding nodes.

When the input data are not assigned to the grid nodes, the spiral search can only locate previously simulated values. The super block search (Journel and

Huijbregts, 1978; Deutsch and Journel, 1998) must then be used to locate the input data. This second search superimposes a coarse grid over the model area, thus creating *super blocks*, and the data inside each block is identified and indexed to that block. The specified search radii are used to construct a template that is centred on the super block containing the point to be estimated. This makes it quick to identify the data inside of the search area. The local data is then exhaustively searched to identify the conditioning data.

The above search routines are only concerned with identifying the most correlated data and ignores their direction. The direction of the data can be taken into account by using the octant search. The octant search divides the surrounding 3D area into eight equal regions. When searching for data, only a maximum number are allowed from each octant. This forces the data to come from different directions at the expense of ignoring closer, but more redundant, data.

5 Kriging

The theory behind SS is based on using every previously simulated value and input data throughout the simulation process (Isaaks, 1990). In practice, only the closest conditioning data are used, up to a maximum number, to keep CPU time reasonable. This assumes the closest data screen the data further away and the additional information from this screened data is deemed small enough to ignore (Isaaks, 1990). The choice of the maximum number is linked to two issues: the speed required to generate a realization, and the accuracy of the kriged estimate and variance.

The impact of the number of data used in kriging on CPU time is controlled by (1) locating the conditioning data, and (2) calculating the kriging weights. For n data, the search is proportional to n, regardless of the search type, and the kriging system calculations are proportional to n^3. So as n increases, the kriging calculations will dominate the CPU time. For example, a 100 x 100 grid was simulated using 300 spatially random data to track the CPU time as n varied from 5 to 300 (Figure 1a). Initially, the change in CPU time is small, but as n increases, the change in the CPU requirement approaches a slope of 3 on a log-log scale.

The uncertainty in kriging is expressed by the kriging variance that is a minimum by construction. Reduction of this variance is only achieved through the addition of more data. Gandin (1963) showed that the change in variance can be bounded when the least informative datum is removed; however, modern computers make the direct calculation of the change quicker than Gandin's method (Zanon, 2004).

For example, 100 conditioning data were randomly chosen and kriging was performed at three arbitrarily chosen test locations (Figure 1b). As n varied from 1 to 100, the kriging estimate and variance were tracked (Figure 2). The best results are achieved when $n = 100$ as indicated by the dotted lines. It is seen that a lower limit of 8 to 10 conditioning data will provide results close to the dotted line, with diminishing returns for $n > 10$.

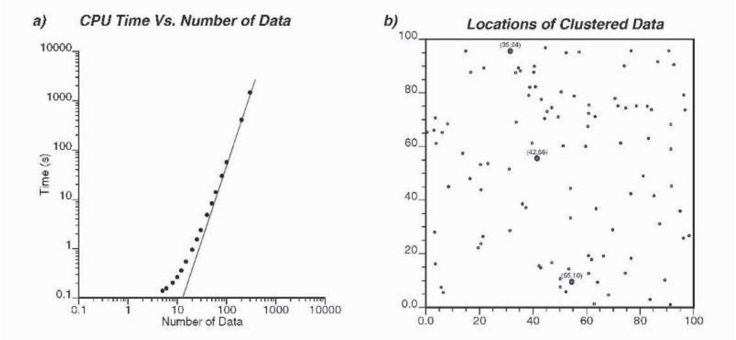


Figure 1. (a) The change in CPU time versus the number of conditioning data. (b)The location of the input data and three test locations (large dots).

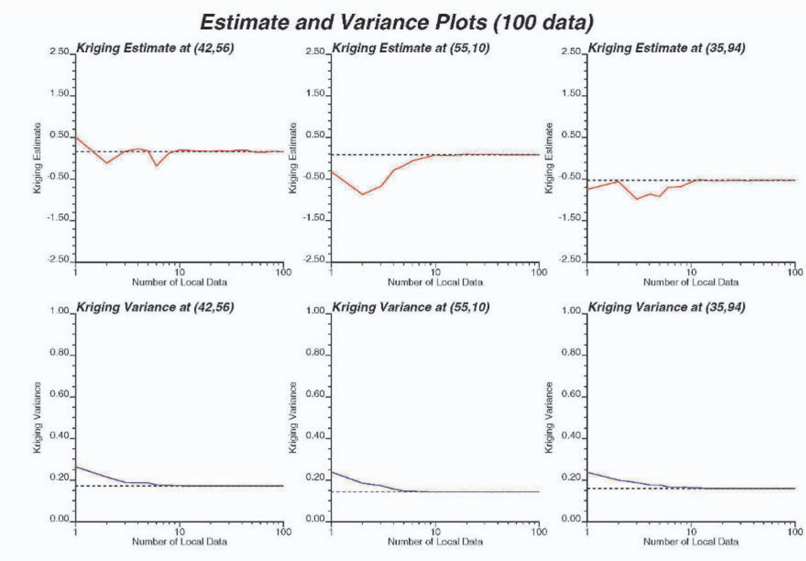


Figure 2. The change in the kriging mean (top) and variance (bottom) for three locations in the area of interest.

6 Final Remarks

Once the modelling process has been completed, the following checks should be performed: reproduction of (1) data values at data locations, (2) the target histogram, (3) the target summary statistics, and (4) the input covariance model. In the multivariate context, this list should also include reproduction of the multivariate distribution and the corresponding summary statistics (Leuangthong, McLennan, and Deutsch, 2004). A visual inspection of the geology can help determine if the model adheres to the expected underlying geological structure.

Failure to satisfy these tests requires some checks and/or changes to the input parameters, this depends on the options available on the software being used. Variance inflation is one cause of poor histogram reproduction. Increasing the number of conditioning data and the search radius, along with the octant search, may help to correct the variance at the cost of increased CPU time. The most common form of poor variogram reproduction is in the long range structure. Using the multiple grid search, along with increasing the number of conditioning data and search radius, can help improve the long range variogram. One general check is to look at the histogram and variogram reproduction of an unconditional simulation. This may help to identify problems caused by the input data and not the program. The assumption of stationarity may be violated and, data permitting, the model should be divided into smaller, more stationary areas.

References

Caers, J., *Adding Local Accuracy to Direct Sequential Simulation*, Mathematical Geology, vol. 32, no. 7, 2000, p. 815-850.

Chilès, J-P., and Delfiner, P., *Geostatistics: Modeling Spatial Uncertainty*, John Wiley & Sons, New York, 1999.

Deutsch, C.V., *Geostatistical Reservoir Modeling*, Oxford University Press, New York, 2002.

Deutsch, C.V. and Journel, A.G., *GSLIB: Geostatistical Software Library and User's Guide, 2nd Edition*, Oxford University Press, New York, 1998.

Gandin, L.S., *Objective Analysis of Meteorological Fields*, Translated from Russian: Israel Program for Scientific Translations (1965), Jerusalem, Israel, 1963.

Gomez-Hernandez, J.J. and Srivastava, R.M., *ISIM3D: an Ansi-C 3 dimentional multiple indicator simulation*, Computer & Geosciences, vol. 16, no. 4, 1993, p. 395-440.

Goovaerts, P., *Geostatistics for Natural Resources Evaluation*, Oxford University Press, New York, 1997.

Isaaks, E.H. and Srivastava, R.M., *An Introduction to Applied Geostatistics*, Oxford University Press, New York, 1989.

Isaaks, E.H., *The Application of Monte Carlo Methods to the Analysis of Spatially Correlated Data*, PhD Thesis, Stanford University, Stanford, CA, 1990.

Johnson, M.E., *Multivariate Statistical Simulation*, John Wiley & Sons, New York, 1987.

Journel, A.G. and Huijbregts, Ch.J., *Mining Geostatistics*, Academic Press, New York, 1978.

Journel, A.G., *Modeling Uncertainty: Some Conceptual Thoughts*, Geostatistics For the Next Century, Kluwer Academic Publications,1993, p. 30-43.

Journel, A.G. and Xu, W., *Posterior identification of histograms conditional to local data*, Mathematical Geology, vol. 26, no. 3, 1994, p. 323-359.

Leuangthong, O., McLennan, J.A., and Deutsch, C.V., *Minimum Acceptance Criteria for Geostatistical Realizations*, Natural Resources Research, accepted April 2004.

McLennan, J.A., *The Effect of the Simulation Path in Sequential Gaussian Simulation*, Centre for Computational Geostatistics Report Four, University of Alberta, Edmonton, Alberta, 2002.

Soares, A., *Direct Sequential Simulation and Cosimulation*, Mathematical Geology, vol. 33, no. 8, 2001, p. 911-926.

Tran, T.T., *Improving Variogram Reproduction on Dense Simulation Grids*, Computers and Geosciences, vol. 20, no. 7/8, 1994, p. 1161-1168.

Verly, G., *The Block Distribution Given a Point Multivariate Normal Distribution*, Geostatistics for Natural Resources Characterization, Part 1, 1984, p. 495-515.

Xu, W., Tran, T.T., Srivastava, R.M. and Journel, A.G., *Integrating Seismic Data in Reservoir Modeling: The Collocated Cokriging Alternative*, Society of Petroleum Engineers, 1992, SPE 24742.

Xu, W. and Journel, A.G., *DSSIM: A General Sequential Simulation Algorithm*, Stanford Center for Reservoir Forecasting Stanford University, 1994.

Zanon, S., *Advanced Aspects of Sequential Gaussian Simulation*, MSc Thesis, University of Alberta, Edmonton, Alberta, 2004.

This page intentionally blank

GEOSTATISTICS BANFF 2004

Volume 2

Edited by

OY LEUANGTHONG

University of Alberta,
Edmonton, Canada

and

CLAYTON V. DEUTSCH

University of Alberta,
Edmonton, Canada

Springer

PETROLEUM

EARLY UNCERTAINTY ASSESSMENT: APPLICATION TO A HYDROCARBON RESERVOIR APPRAISAL

GUILLAUME CAUMON[1,2] and ANDRÉ G. JOURNEL[1]
[1] *Petroleum Engineering Dept., Stanford University*
[2] *ENS Géologie, INPL – CRPG, Nancy, France*

Abstract.

Assessment of uncertainty of global resources from sparse appraisal data is a difficult challenge. While many algorithms have been defined to compute one single "best" estimate $a^\star$ of the unknown global value a, assessing the uncertainty calls for the definition, necessarily subjective, of a randomization process.

Most error assessment algorithms, including bootstrap resampling, consider a randomization of the global estimate $a^\star$. We suggest a joint randomization of both the unknown a and its estimate $a^\star$ within a Bayesian framework, given alternative plausible geological scenarios. This allows for:

- considering a prior probability distribution for the unknown target value a based on analog studies,
- obtaining the data likelihood by spatial bootstrap instead of using some arbitrary analytical distribution,
- assessing the value of data in reducing the prior uncertainty, a prerequisite to decide on new data acquisition strategies.

The proposed procedure does not call for data independence nor Gaussian assumptions which are seldom met in practice. It accounts explicitly for alternative geological interpretations of the quantitative data available, a critical source of uncertainty too often ignored. The method is applied to a complex synthetic fluvial reservoir.

1 Introduction

Notwithstanding shortage of data, assessing the uncertainty about a global attribute value is difficult: such a global attribute is inherently unique, as opposed a local estimation error for which replicates over the same study field can be found (Journel and Huijbregts, 1978). Each reservoir is unique, and it is no trivial task to conceive a probability distribution for any of its global or average attributes.

The definition of a confidence or a probability interval for any unknown attribute value a requires some kind of randomization of the estimated value $a^\star$, the

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 551–557.

unknown attribute a itself, or both a and $a^\star$; by "randomization", we understand a process by which replicates are defined and their probability distributions evaluated. Two classes of uncertainty assessments of global properties are prevalent, see (Biver *et al.*, 1996).

Distribution for the global estimate. The Bootstrap algorithm (Efron, 1979) repeats the global estimation procedure on alternative data sets resampled with replacement from the actually observed data. This approach considers the well data values to be independent one from another, which is never true in practice. Solow (1985) adds spatial dependency specified by a covariance matrix to the bootstrap.

Haas and Formery (2002) use an analytical expression of Efron's bootstrap approach: the set of independent facies samples $\mathbf{n}$ given a particular facies proportion $\mathbf{x}$ follows a multinomial distribution $p_{\mathbf{x}}(\mathbf{n})$; the authors show by Bayesian inversion that the distribution $p_{\mathbf{n}}(\mathbf{x})$ of proportions given the samples is Dirichlet, assuming a prior distribution $p(\mathbf{x})$ of facies proportions either uniform or Dirichlet. Haas and Formery (2002) further propose to modify the parameters of the posterior Dirichlet distribution $p_{\mathbf{n}}(\mathbf{x})$ to account for data redundancy as evaluated by indicator kriging; this is questionable since the multinomial distribution $p_{\mathbf{x}}(\mathbf{n})$ used in the analytical development is not multinomial if the samples $\mathbf{n}$ display spatial dependence.

Moreover, the bootstrap and the Dirichlet distribution cannot easily account for preferentially located data and incorporate local secondary information as provided by seismic data.

In the spatial bootstrap method (Journel, 1993; Norris *et al.*, 1993), alternative sets of data are resampled from whole simulated fields. This resampling method accounts for any prior model of spatial dependency between the data, and allows for integration of secondary information. Recently, Journel and Bitanov (2004) have applied the spatial bootstrap for assessing the uncertainty of net-to-gross estimates from early exploration data.

Distribution for the unknown true value. Bayesian approaches start with a prior probability distribution for the unknown global value A, which is then updated by a data likelihood typically assumed Gaussian. This allows incorporating one's expertise into the uncertainty assessment through the prior distribution. The main limitation of Bayesian methods lies in the specification of a realistic data likelihood: traditionally used Gaussian models fail to recognize the typical complexity of the relations between data and between the data and the true value, e.g., non-linearity and heteroscedasticity.

2 Global uncertainty assessment

We now propose a workflow which reconciles the two previous approaches, overcoming their respective drawbacks. This workflow considers three jointly related

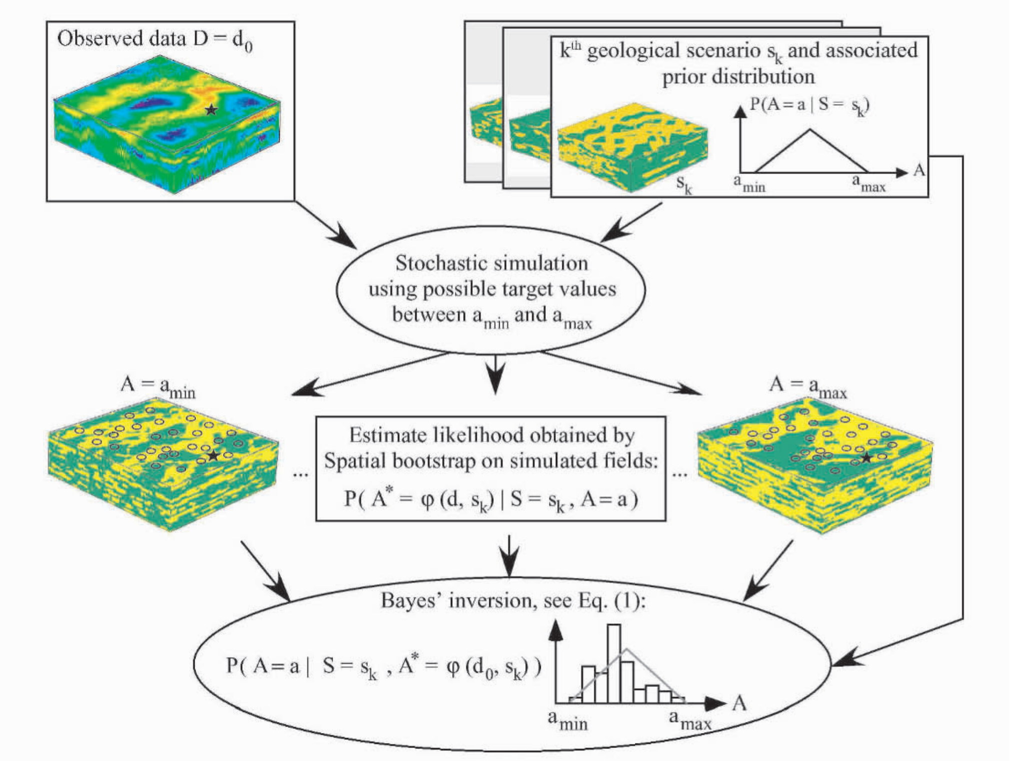

Observed data D = d$_0$
kth geological scenario s$_k$ and associated prior distribution
P(A = a | S = s$_k$)
s$_k$
a$_{min}$
a$_{max}$
A
Stochastic simulation using possible target values between a$_{min}$ and a$_{max}$
A = a$_{min}$
A = a$_{max}$
Estimate likelihood obtained by Spatial bootstrap on simulated fields:
P(A* = φ (d, s$_k$) | S = s$_k$, A = a)
...
...
Bayes' inversion, see Eq. (1):
P(A = a | S = s$_k$, A* = φ (d$_0$, s$_k$))
a$_{min}$
a$_{max}$
A

Figure 1. Flow chart of the proposed procedure, see text for details. Initial well location is figured by a star. Circles represent alternative well locations.

sources of uncertainty and the corresponding random variables denoted with capital letters: S for the geological scenario, A for the true global value, $A^\star$ for its estimate.

Geological scenario S. The lack of subsurface data leaves room for various possible interpretations of the geological setting of the reservoir under study. We believe the sedimentological and structural interpretations to be the major sources of uncertainty and propose to account for it through a set of deemed plausible geological scenarios $\{s_1, \ldots, s_K\}$, considered as the possible outcomes of a discrete random variable S. A scenario s_k could reduce to a mere variogram model, or, better, consist of a prior conceptual image of spatial patterns as depicted by a training image (Fig. 1, 1^{st} row). We suggest to use training images and the corresponding multiple point geostatistical formalism, since these can carry geological interpretation more comprehensively than variograms (Guardiano and Srivastava, 1993; Strebelle, 2002).

The estimator $A^\star$. The observed data set $\mathbf{d}_0$ is processed into an estimated value $A^\star = a^\star$ of the unknown global value a. That estimation process can be represented by a function $\varphi(\mathbf{d}, s_k)$ mapping any particular set of data $\mathbf{d}$ into some "best" estimate $A^\star = a^\star$ under a geological scenario s_k. Typically in hydrocarbon application, the function $\varphi(\mathbf{d}, s_k)$ relies on a seismic-to-well calibration, (Fournier and Derain, 1995; Caers and Ma, 2002; Journel and Bitanov, 2004).

One way to randomize the estimate $a^\star$ is to randomize the data $\mathbf{d}$ themselves. Therefore, we will use a probabilistic notation (capital letter) for the data event, with $\mathbf{D} = \mathbf{d}_0$ denoting the actual data values observed. In our case, these data comprise both seismic impedance and well observations. $\mathbf{D} = \mathbf{d}$ denotes any alternative data set, where some aspect of the data is changed. In this paper, we use spatial bootstrap (Journel, 1993; Norris *et al.*, 1993) to obtain such alternative data events: for a stochastically simulated field of known global value a, well locations hence well values can be randomized under some drilling constraints (Fig. 1, 3^{rd} row). For each new set of well locations, the estimation procedure can be repeated, providing the likelihood probability $P(A^\star = a^\star | A = a, S = s_k)$ for the estimate $a^\star$ to occur, given the true value a and the geological scenario s_k.

Ideally, randomizing the estimate $a^\star$ into the random variable $A^\star$ would call for randomizing both the data $\mathbf{d}_0$ into $\mathbf{D}$ but also the estimation algorithm $\varphi(\cdot)$. In this paper, the estimation algorithm $\varphi(\cdot)$, the seismic data and the well drilling strategy are frozen.

The unknown value A. Most error assessment algorithms, including bootstrap resampling (Efron, 1979), consider only a randomization $A^\star$ of the estimate $a^\star$. We suggest that the unknown global value a should be simultaneously randomized into a random variable A, conditionally to each geological scenario $S = s_k$, $k = 1, \ldots, K$ (Fig. 1, 2^{nd} row).

Selecting a prior probability $P(A = a)$ for the random variable A to take any value a is a delicate task, since that prior probability should almost never be a uniform distribution. Instead, we consider a different prior probability distribution $P(A = a|S = s_k)$ specific to each geological scenario $s_k, k = 1, \ldots, K$. Such a distribution can be for instance obtained from sedimentological analogs.

BAYESIAN INVERSION

For each geological scenario s_k, the probability distribution $P(A = a|S = s_k)$ for the unknown attribute A to take any value a can be updated by the "best" estimate retained $a^\star = \varphi(\mathbf{d}_0, s_k)$ using a Bayesian inversion (Fig. 1, 4^{rd} row):

$$P(A = a|S = s_k, A^\star = a^\star) \; = \; \frac{P(A^\star = a^\star|S = s_k, A = a) \cdot P(A = a|S = s_k)}{P(A^\star = a^\star|S = s_k)} \qquad (1)$$

3 Application: net-to-gross uncertainty in a channelized reservoir.

The proposed methodology is now applied to the exhaustively known Stanford V fluvial reservoir (Mao and Journel, 1999). To mimic an actual appraisal situation, two preferentially located wells and an impedance cube have been extracted from the reference reservoir model. The NTG observed along the wells is .58, a value higher than the true and unknown reservoir NTG 0.48, because the two appraisal wells are located in a low impedance area. The goal of the study is to assess from this data set $\mathbf{d}_0$ the uncertainty of the reservoir net-to-gross (NTG) defined as the proportion of sand.

3.1 CHOICE OF GEOLOGICAL SCENARIOS

Two geological scenarios have been retained for this study, both corresponding to a channelized reservoir. The first scenario s_1 is represented by a conceptual training image with a NTG value of 0.46, see Figure 2-B. The corresponding prior probability distribution $P(A = a|S = s_1)$ was chosen to be triangular defined between 0.24 and 0.68 with a mode of 0.46.

The second scenario, deemed "pessimistic", assumes a lower channel density. This scenario s_2 is represented by a training image having a NTG of 0.35. The corresponding prior distribution $P(A = a|S = s_2)$ is a triangular between 0.13 and 0.57, with a mode of 0.35.

3.2 SPATIAL BOOTSTRAP

For the geological scenario s_1, 20 classes $\{a_1, \ldots, a_{20}\}$ of possible NTG values were considered. Two conditional stochastic realizations were generated for each of these NTG classes using the snesim algorithm (Strebelle, 2002). Within the area of 20% lowest vertically averaged impedance values, 600 alternative sets of two wells were resampled from each of these $2 \times 20 = 40$ simulated reservoirs. An estimation based on a co-located Bayesian seismic-to-well calibration (Journel and Bitanov, 2004) was applied to these $20 \times 2 \times 600 = 24,000$ alternative data sets to obtain 20 likelihoods $P(A^\star = a^\star|A = a_m, S = s_1)$ used on the right-hand side of Equation (1). This procedure was repeated for scenario s_2.

3.3 PROBABILITY INTERVALS OBTAINED

The specific estimation method by Journel and Bitanov (2004) applied to the appraisal data set $\mathbf{d}_0$ under scenarios s_1 and s_2 yields NTG estimates of 0.43 and 0.36, respectively. The Bayesian inversion (1) produces the final probability distributions (bar charts) displayed in Figures 2-C,D.

Under the geological scenario s_1, the procedure reduces the prior uncertainty, but does not entail any shift of the distribution. Under scenario s_2, the distribution is shifted towards the higher NTG values. This is explained by the relative concordance between the scenario s_1 (NTG $\simeq$ 0.46) and the actual Stanford V

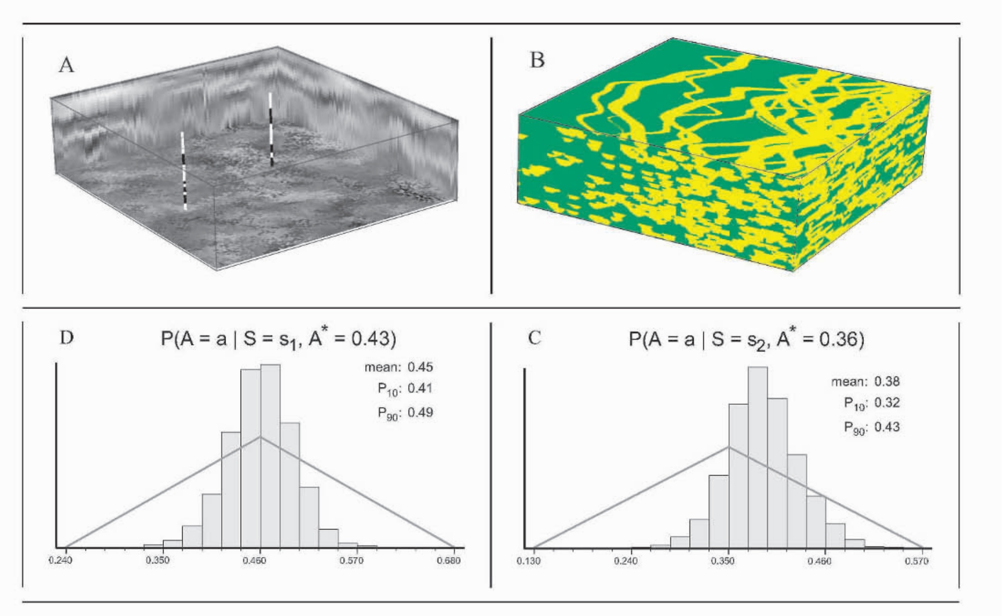


Figure 2. A- Appraisal data set used for the study. B- Training image corresponding to scenario s_1. C,D- Results obtained with scenarios s_1 and s_2.

reservoir 0.48 NTG, while the scenario s_2 (NTG $\simeq$ 0.35) is clearly pessimistic yet with a high appraisal wells' NTG of 0.58.

4 Conclusion

As opposed to more traditional approaches, the proposed workflow (1) does not call for independence nor Gaussian assumptions which are rarely met in practice, and (2) considers an uncertainty model based on the joint randomization of both the true global value and its estimate. This approach is not specific to a particular multiple point or variogram-based simulation algorithm; this leaves room for the practitioner to choose the simulation method best suited to the problem at hand. Modeling of non stationary variables is achieved as far as permitted by the retained simulation method (e.g., its ability to account for vertical proportion curves or 3D proportion cubes).

The Bayesian updating suggested calls for a prior distribution for the global unknown variable, one which can be obtained from analog studies. The data likelihood is obtained by spatial bootstrap, which accounts for geological interpretation and the consequent model of data dependence. Such spatial bootstrap allows assessing the value of quantitative data in reducing prior uncertainty, a prerequisite to decide about new data acquisition strategies.

As with any Bayesian approach, the proposed workflow depends heavily on the prior probability distribution retained for the true target value, more so as the

This page intentionally blank

observed data are fewer and less informative. The consideration of various alternative geological interpretations addresses directly one main source of uncertainty during appraisal stage; determining the prior probability attached to each scenario is a difficult and still open problem. Uncertainty in the data themselves (coming from well logs interpretations and seismic data processing) could also be taken into account when determining the data likelihood.

Acknowledgements

We would like to thank ChevronTexaco ETC and the other affiliates of the Stanford Center for Reservoir Forecasting for their support. Thanks to Sébastien Strébelle and Jef Caers for their input on this work, and to the anonymous reviewers for their constructive remarks.

References

P. Biver, P. F. Mostad, and A. Guillou. An overview of different techniques to quantify uncertainties on global statistics for geostatistical modeling. In E. Y. Baafi and N. A. Schofield, editors, *Geostatistics Wollongong '96*, volume 1, pages 573–584. Kluwer, Dordrecht, 1996.

J. Caers and X. Ma. Modeling conditional distributions of facies from seismic using neural nets. *Mathematical Geology*, 34(2):139–163, 2002.

B. Efron. Bootstrap methods: another look at the jackknife. *Annals of Statistics*, 7:1–26, 1979.

F. Fournier and J. F. Derain. A statistical methodology for deriving reservoir properties from seismic data. *Geophysics*, 60(5):1437–1450, 1995.

F. B. Guardiano and R. M. Srivastava. *Multivariate Geostatistics: Beyond Bivariate Moments*, volume 1, pages 133–144. Kluwer, Dordrecht, 1993.

A. Haas and P. Formery. Uncertainties in facies proportion estimation, I. theoretical framework: the Dirichlet distibution. *Math. Geol.*, 34(6):679–702, 2002.

A. G. Journel and A. Bitanov. Uncertainty in n/g ratio in early reservoir development. *Journal of Petroleum Science and Engineering*, 42(xx (in press)), 2004.

A. G. Journel and C. Huijbregts. *Mining Geostatistics.* Academic Press, NY; Blackburn Press, NJ (reprint, 2004), 1978.

A. G. Journel. Resampling from stochastic simulations. *Environmental and Ecological Statistics*, 1:63–83, 1993.

S. Mao and A. G. Journel. Conditional 3d simulation of lithofacies with 2d seismic data. *Computers and Geosciences*, 25(7):845–862, 1999.

R. Norris, G. Massonat, and F. Alabert. Early quantification of uncertainty in the estimation of oil-in-place in a trubidite reservoir. In *SPE Annual Technical Conference and Exhibition (SPE 26490)*, 1993.

A. R. Solow. Bootstrapping correlated data. *Mathematical Geology*, 17:769–775, 1985.

S. Strebelle. Conditional simulation of complex geological structures using multiple-point statistics. *Math. Geol.*, 34(1), 2002.

RESERVOIR FACIES MODELLING: NEW ADVANCES IN MPS

ANDREW HARDING, SEBASTIEN STREBELLE, MARJORIE LEVY, JULIAN THORNE, DEYI XIE, SEBASTIAN LEIGH AND RACHEL PREECE
ChevronTexaco Energy Technology Company, PO Box 6046, San Ramon, CA 94583-0746, USA.
ROBERT SCAMMAN
ChevronTexaco Overseas Petroleum, PO Box 430, Bellaire, TX 77402-0430. USA

Abstract

Over the last several years, Multiple Point statistical Simulation (MPS) has emerged as a practical tool in the characterization and modeling of petroleum reservoirs. In this paper, we describe recent developments in facies modeling at ChevronTexaco and illustrate this with a case history of its use at a ChevronTexaco operated field. Our approach is based on our implementation and continuing development of MPS. Our MPS workflow models depositional facies in a four step process.

The first step is the construction of the Training Image used by MPS algorithms. The second step is the compilation of a three-dimensional azimuth field. This information field imparts areal geological trends at the scale of the reservoir model and is derived from well and seismic information together with sub-regional geological trends. The third step is to calculate a facies probability cube. The cube defines the facies proportions in the model and is derived from well and seismic information. The final step is the application of the MPS. The result is a realization of the facies model with all the characteristics of the three components described. The workflow facilitates construction of multiple realizations of the model.

Uncertainty in facies occurrence is not, however, limited to multiple realizations. Provided that the training image is sufficiently rich with a sound knowledge of the geology, correctly represented, a single training image can be use and multiple azimuth fields and facies probability scenarios incorporated into the modeling. This latter technique will be illustrated.

Our approach results in models that represent geology very well. Obtaining this characteristic is not solely a function of the algorithms and workflows used. Sound and detailed sequence stratigraphic analysis is an absolute prerequisite, even where well data is sparse.

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 559–568.

1 Introduction

The construction of hydrocarbon-reservoir geocellular models for flow simulation requires the integration of diverse disciplines in geophysics, stratigraphic geology and geostatistics. Commercially available modeling programs, in general, perform an excellent job of modeling continuous properties, such as porosity and permeability, but modeling categorical properties such as geologic facies presents more challenging issues. Practitioners debate the most appropriate geologic facies to model, with depositional facies being favoured by the stratigraphers and lithofacies (or electrofacies or petrofacies, depending on terminological details) by petrophysicists. Further complications arise if seismic data are used to condition the model.

Multiple-Point geostatistical Simulation (MPS) offers both the facies geometry realism of object-based Boolean modeling methods and the seismic and well conditioning capability of indicator simulation methods. ChevronTexaco has implemented the MPS modelling code developed by the Stanford Center for Reservoir Forecasting as a proprietary plug-in to the Gocad environment, together with numerous algorithmic improvements

The need for some sort of facies division in hydrocarbon reservoir geocellular modelling is generally accepted and has been discussed by many authors, particularly Deutsch and Journel (1993). In summary, different geological facies have different spatial statistics and petrophysical properties and these need to be honoured in the modelling process.

2 Depositional Facies

In this paper, we discuss the use of Multiple-Point geostatistical Simulation to model depositional geological facies. We define the latter following Walker (1992) as “a body of rock characterized by a particular combination of lithology, physical and biological structures that bestow an aspect different from the bodies of rock above, below and laterally adjacent.” The depositional facies in each well section is determined by expert geologic analysis of well data. Ideally, core data are required, but these are never comprehensive enough for complete characterization in oil-field cases. Practically, geologists use a combination of core and wireline data and supplement this with outcrop and other analogue information: we follow this methodology herein. Modern well-log suites that include micro-resistivity imaging tools can give us resolution approaching that of core data, when properly calibrated, but these were not available for this study.

The main advantage of using depositional facies for reservoir modeling over the other types of rock facies is that we often have a good knowledge of geometry and spatial relationships of the depositional facies interpreted in wells derived from outcrop and sub-surface analogues; with multiple-point geostatistical methods we can represent these in our models.

The main disadvantage in using depositional facies for modeling is that the reservoir properties inside a particular facies volume are not well defined; there is a significant uncertainty in porosity and permeability. Our workflow addresses this uncertainty directly in the modelling of the continuous properties but that is beyond the scope of this paper.

Previous workers have used indicator simulation or Boolean simulation techniques to model depositional facies away from well control. The advantages and disadvantages of these techniques have been discussed elsewhere and do not require further debate here.

3 Multiple Point Geostatistical Modeling and Simulation (MPS)

The theory and the mechanics of the MPS method have been described elsewhere (Guardiano and Srivastava, 1993; Strebelle and Journel, 2001; Strebelle, 2002). Other workers (Strebelle, Payrazyan and Caers, 2002; Liu et al, 2004) have described its application to binary and three-fold facies subdivisions.

The MPS method uses a three-dimensional training image to derive the probability of a particular facies occurrence in a modelling cell and this probability information is used in a statistical simulation of the facies occurrence in the modelling grid.

The training image is a three-dimensional conceptual model of the depositional facies with their correct shapes and areal associations; it captures complex spatial relationships between multiple facies, and non-linear shapes such as sinuous channels. It is analogous to the traditional geological block diagram often used to give a 3D rendering of the geology. The training image contains no absolute location information and it defines the appropriate relative scales present in the model (including areal extent, facies body thicknesses and approximate facies proportions). It is compiled by workers' knowledge of the depositional environment and draws heavily on outcrop and subsurface analogues. It does not require a tie to well data. Object-based modeling techniques are well suited to construct the training image cube. When modelling within the same geological environment, only one training image is required, containing a multiplicity of facies and sufficiently rich in the variability of these facies; sufficient variation of facies dimensions and associations must be present in the training image to represent the possible ranges interpreted to be present in the geology. If our knowledge of the geology is limited such that a wider range of uncertainty in facies dimensions is present, then multiple training images may be required and would be treated as an uncertainty parameter. In this example we used a single training image.

The facies and the facies associations defined by the training image can occur anywhere in the model. This is the condition that the multiple-point process simulation is stationary. To modify this condition of stationarity, further probability control is required in the multiple-point statistical simulation and we have developed a four-stage workflow to achieve this (Figure 1). The final probability field used for the facies simulation is generated by a combination of the training image, an azimuth field, which modifies the geometry of the training image to conform to local geology, and a facies probability cube, which provides soft control on the local facies proportions. The facies probability cube is derived from our geological interpretation of the available data including well logs, seismic and regional geological knowledge.

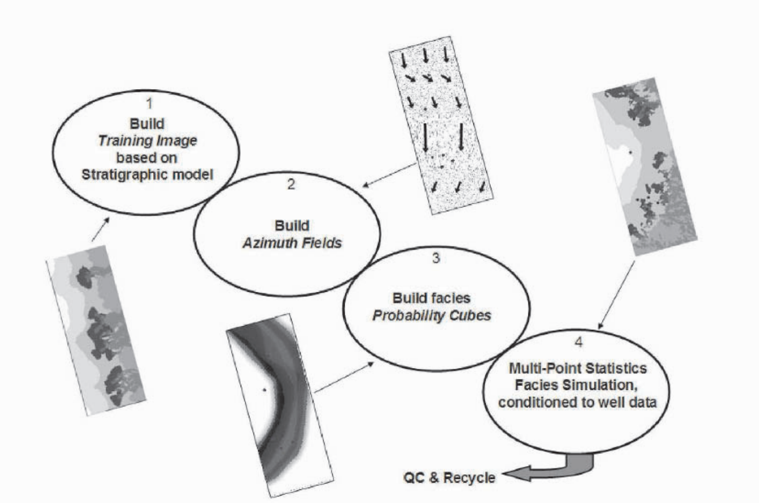


Figure 1: Workflow diagram

We will describe the application of this workflow to a case history.

4 Case History Geological Setting

Our example is a reservoir model built by ChevronTexaco during 2003 for the purposes a field development planning. The field is located offshore West Africa. The Cretaceous-age reservoir was deposited in a transitional setting with terrestrial-to-marine paleo-environments succeeding each other, controlled by sea-level changes. Sediment supply was highly variable, resulting in mixed clastic and carbonate lithology. The complexity of this reservoir has presented a major challenge to geological modelling: a clear understanding of the stratigraphy and depositional facies interpretation is required before any modelling can begin.

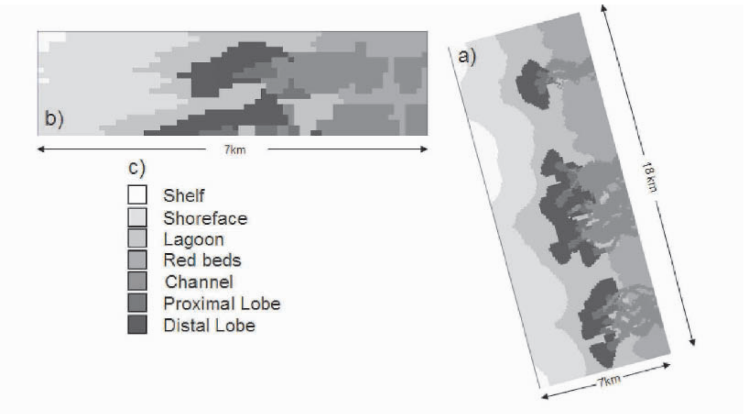


Figure 2: a) Map view of Training Image; b) Section view of Training Image, flattened top and base; c) Legend for Depositional Facies

Figures 2 and 3 show the training image we have constructed, using simple three-dimensional-object building blocks. The map view in Figure 2 shows the basic areal facies architecture with terrestrial environments (red-beds) on the right and marine environments (shoreface and shelf) on the left with a lagoonal environment transitional between the two. Seasonal channels cut through the red-bed sequence to deposit clastic sediments both in the channels themselves and in the lagoon as proximal and distal lobes. Each of these depositional facies can be broken down into separate sub-groups but that would result in complexity too great for practical modelling. The hinterland beyond the right-hand extreme of the model was mostly arid and periods of drought allowed the build-up of carbonates in the lagoon and the deposition of carbonate sands and muds in the offshore. The variability of reservoir properties was taken into account as an uncertainty in the continuous property modelling stage using discrete scenarios.

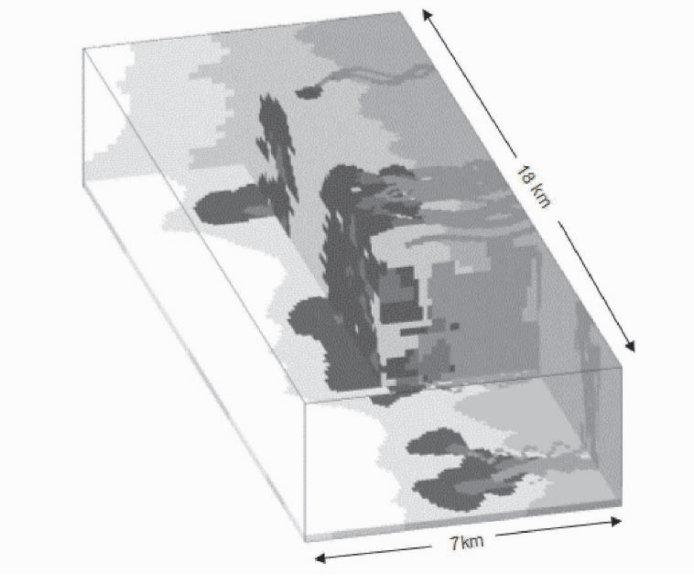


Figure 3: Cut-away block diagram view of flattened training image showing the depositional facies distribution. See Figure 2 for facies legend.

To construct the training image, we first populate a 3-dimensional grid with facies belts such as the red-beds, lagoon, shoreface and shelf. We then use an object simulation technique to add objects such as channels and attach lobes to the distal end of the channels; rules for facies associations are imposed to govern the superposition of facies and the co-existence of facies; for example, channels cut through red-beds and lagoon but do not reach the shoreface and shelf; distal lobes are present at the channel mouth and these lobes prograde into the lagoon but are not associated with red-beds or shoreface. Deposition facies dimensions (such as channel widths and thicknesses, channel amalgamation, width of the shoreface etc.) are contained in the training image, which must also have sufficient variability in facies dimensions to cover the range interpreted to be present in the sub-surface; but we are also constrained by the need to keep the training image from growing too large: this can increase simulation run-times, which are dependent on building a search-tree to determine multiple-point facies probabilities.

The training image was built without any assumption about its orientation in space (except for the sense of the vertical direction). This allows us to use the same training image for all the reservoir zones, as we interpret all zones to have the same geological model, but as explained below, we will incorporate differing facies proportions. Our entire model consisted of 15 reservoir zones defined within sequence-stratigraphic flooding surfaces and sequence boundaries; a unique stratigraphic interpretation was made for each zone of the model.

5 Application of the Workflow

The workflow is illustrated in Figure 1 and is implemented in ChevronTexaco's version of the Earth Decision Science's Gocad visualization and modelling software. We have described the construction of the training image above. The model is constructed in a three-dimensional stratigraphic grid which conforms to the interpreted stratal surfaces defining the model. These stratal surfaces are generated from depth-converted seismic horizons with proportional layering between.

The second stage is the construction of the azimuth field, which imparts the sense of the spatial orientation of the training image with respect to the model: the azimuth field effectively twists the training image to allow for local variations of depositional strike. The field can be fully 3D, but in our example we do not have that degree of resolution; rather, we made a facies map interpretation for each reservoir interval using well and seismic information together with sub-regional geological trends. We have constructed a two-dimensional azimuth field for each reservoir zone.

Figure 4 shows the process to construct the azimuth field for each zone. First, lines of constant azimuth are drawn using the geologist's facies map as a guide. Standard Gocad functions are used to convert the information from a vector field to a property on the stratigraphic grid used for modelling. The resulting azimuth field is not a unique interpretation and this uncertainty could be modelled by generating several versions of the azimuth map but for simplicity, we have only worked with a single scenario for each zone.

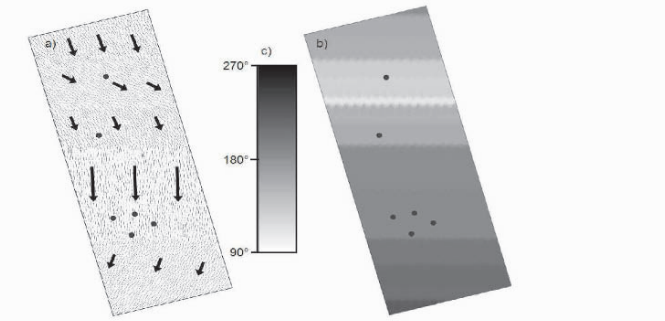


Figure 4: Azimuth Field for one grid layer (map view with same dimensions as figure 2). a) Vector field derived from lines of constant azimuth; b) Grey-scale representation of azimuth property on modeling grid; c) grey scale for b). Well locations at the top of the model are shown as circles.

The third stage is the construction of a facies probability cube based on a ChevronTexaco proprietary method. The inputs to this process are as follows. For each stratigraphic interval studied, we construct a vertical facies proportion curve. This defines the facies proportion for each layer of grid cells and is based partly on well information and partly on geological interpretation. We also provide a depocenter map to define where the facies are deposited in that interval. A proprietary algorithm combines this information to generate a cube of facies probabilities; in our example, we have seven facies and a probability value for each is required in every grid cell. Figure 5 shows examples of a horizontal slice through the probability cube showing the values for the seven depositional facies. These probability values impose a strong constraint on the multi-point simulation in its early stages. For example, the shelf facies is constrained to exist only in a restricted region in the west of the model. The facies probability changes according to the different facies distribution interpretations for each reservoir zone. For example, the channel and lobe facies are not present in some reservoir intervals and the corresponding facies probabilities are set to zero to facilitate this. The facies probability cube has to be consistent with the azimuth information; they are not independent quantities: the azimuth provides the correct orientation of the facies depositional trends and the probability information determines the correct facies proportions in the model.

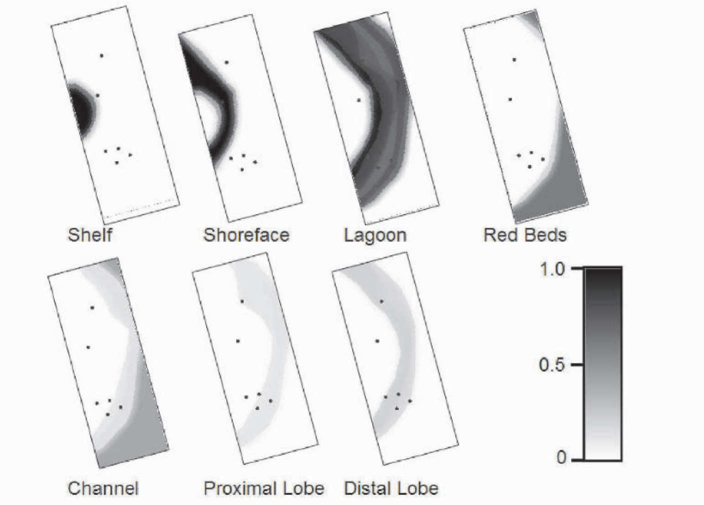


Figure 5: Facies Probability Cube for single grid layer (map view with same dimensions as figure 2) and grey-scale. Well locations at the top of the model are shown as circles.

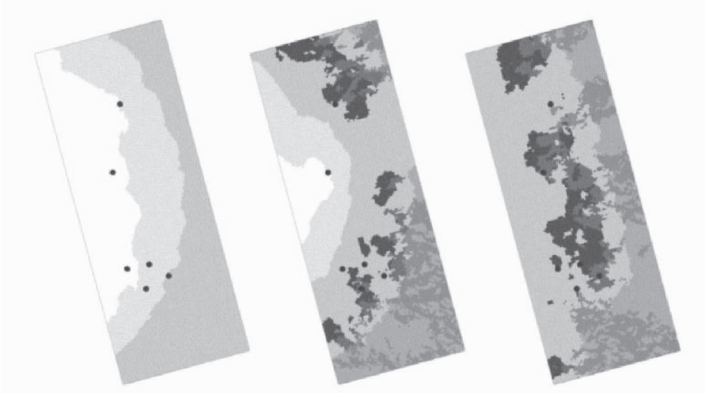


Figure 6: Middle case facies (MPS) realizations (map view with same dimensions as figure 2) for three model layers. Centre panel grid layer corresponds to that shown in figures 4 and 5; see figure 2 for facies legend. Well locations at the top of the model shown as circles.

The final stage of the workflow is the multiple-point geostatistical simulation. This process combines the probability information from the facies probability cube, the training image and facies interpreted in the wells to produce a simulation. Figure 6 shows three layers in different reservoir zones of the final model. The impact of the facies probability cube is clearly demonstrated by the different facies proportions realized: the shelf and shoreface facies are absent in the third layer. Multiple realizations with different seeds are possible; therefore, the results are reviewed carefully to ensure that our geological ideas are correctly represented in the resulting model. Recycling through the workflow may be required to adjust facies proportions. The idealized facies shapes of the training image are considerably modified in the simulation process but the overall facies geometry and associations are preserved. The same training image is used in all 15 reservoir zones.

The training image is a very powerful tool that, with the multiple-point geostatistical simulation, can allow the construction of realistic geological models. This process is highly dependent, however, on the use of a correct depositional model of the geology.

6 Incorporation of Facies Proportion Uncertainty

It is standard practise within ChevronTexaco to make multiple geocellular models for our reservoirs to capture uncertainty in the geology and reservoir properties. In this paper, we illustrate the process to capture the uncertainty in facies proportions.

The sparse well control (Figure 5) does not allow us to define the facies distribution at the level of detail required for flow simulation. We have therefore postulated three scenarios to represent low, middle and high cases of reservoir occurrence. The reservoir quality rocks are contained within the shoreface, channel and lobe facies. For each reservoir zone, three facies proportion scenarios were constructed using facies probability cubes. The criteria for differentiating these were based on our geological knowledge of similar fields nearby and analogous outcrops. This is a subjective process and the rigorous determination of the correct assignment of 10%, 50% and 90%

cumulative probability thresholds is beyond the scope of this study and may not be possible.

Examples of the three cases of reservoir occurrence for one reservoir layer are shown in Figure 7, all based on the same training image. All three realizations have the same results at the well locations as the facies data at the wells are honoured explicitly in the MPS process. In the northern well, for both low and middle cases, a distal lobe approaches the well location closely, whereas the high case shows the location to be clearly separated from a lobe. This can be explained by considering the three-dimensional nature of the simulation.

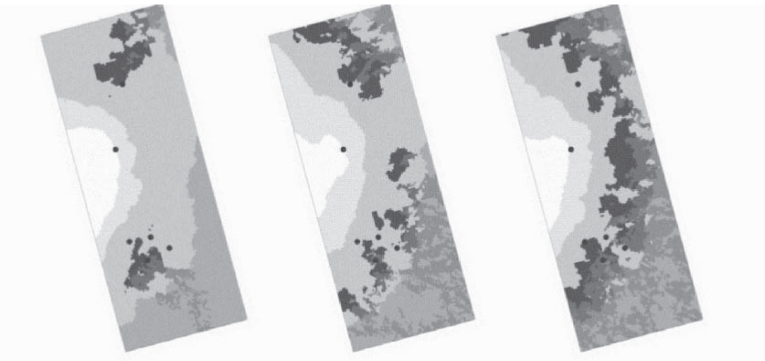


Figure 7: Reservoir Proportion Uncertainty: low, middle and high cases for reservoir facies uncertainty for one reservoir layer (middle example in Figure 6). Map view with same dimensions as figure 2; see figure 2 for facies legend; well locations at the top of the model shown as circles.

7 Conclusion

This paper has illustrated a real case example of the use of Multiple Point geostatistical Simulation to model complex reservoir facies. In summary: 1) The MPS algorithm has been improved to permit a single training image, which is constructed for a particular geological environment to be used to model multiple stratigraphic zones. 2) A relatively large number of facies have been employed, which has enabled the complexities of the reservoir to be well represented in the model. 3) A workflow has been used that allows the integration of geological facies' geometry, associations and heterogeneity with varying azimuth and facies proportions and 4) the workflow allows distinctly different geological scenarios to be modelled, permitting an improved understanding of the impact of uncertainty in facies distribution on the reservoir continuity and pore volume. The MPS algorithm is implemented in ChevronTexaco's version of the Earth Decision Science's Gocad visualization and modelling software and is being use extensively throughout ChevronTexaco's earth science community.

The MPS method results in models that represent geology very well. Obtaining this characteristic is not solely a function of the algorithms and workflows used. Sound and detailed sequence-stratigraphic analysis is an absolute prerequisite, especially where well data are sparse.

8 Acknowledgements

The authors wish to thank Sonangol, Cabinda Gulf Oil Company, Agip Angola Production B.V. and Elf Petroleum Angola for permission to publish this paper. We received generous support from the management of ChevronTexaco Overseas Petroleum Inc., and ChevronTexaco Energy Technology. Special thanks go to Bryan Bracken and Jennifer Ayers of ChevronTexaco for their work in developing the depositional facies model used and numerous discussions on its application to this study.

References

Deutsch, C., and A. Journel, 1993, Combining knowledge from diverse sources: an alternative to traditional data independence hypotheses: Mathematical Geology, v. 25, no. 3, p. 329-355.

Guardiano, F. and Srivastava, R.M. (1993). Multivariate Geostatistics: Beyond Bivariate Moments. In Soares, A., editor, *Geostatistics-Troia*, v. 1, p. 133-144. Kluwer Academic Publications, Dordrecht.

Journel, A., 2002, Combining knowledge from diverse sources: an alternative to traditional data independence hypotheses: Mathematical Geology, v. 34, no. 5, p. 573-596.

Liu, Y., A. Harding, W. Abriel, and S. Strebelle, 2004, Multiple-Point Simulation Integrating Wells, Three-dimensional Seismic Data, and Geology: AAPG Bulletin, v. 88, no. 7, p. 905-921.

Strebelle, S., 2002, Conditional Simulation of Complex Geological Structures Using Multiple-Point Statistics: Mathematical Geology, v. 34, no. 1, p. 1-22.

Strebelle, S., and Journel, A.: "Reservoir Modeling Using Multiple-point Statistics," paper SPE 71324 presented at the 2001 SPE Annual Technical Conference and Exhibition, New Orleans, Sept. 30-Oct. 3.

Strebelle, S., K. Payrazyan, and J. Caers, 2002, Modeling of a Deepwater Turbidite Reservoir Conditional to Seismic Data Using Multiple-Point Geostatistics: paper SPE 77425 presented at the 2002 SPE Annual Technical Conference and Exhibition, San Antonio, Sept.29-Oct. 2.

Walker. 1992: Facies, Facies Models and Modern Stratigraphic Concepts in Facies Models: Response to Sea Level Change, Geological Association of Canada, Edited by Roger G Walker and Noel P James, ISBN 0-919216-49-8.

FITTING THE BOOLEAN PARAMETERS IN A NON-STATIONARY CASE

H. BEUCHER, M. BENITO GARCÍA-MORALES AND F. GEFFROY
Ecole des Mines de Paris, Centre de Géostatistique, Fontainebleau, France

Abstract.

Classically, the non stationary lithofacies distribution inside reservoirs is described by the 3D distribution of their proportions. This approach is very attractive because proportions have a physical meaning. Moreover their 3D distribution reflects the qualitative information coming from geology or the quantitative constraints derived from seismic attributes. In models such as the truncated gaussian, the proportions are directly used in the simulation process. In object-based models, such as the Boolean model, the problem is more complex because the proportions are the results of two sets of parameters: the object description (shape and dimensions) and their 3D distribution.

The non stationarity in an object-based method can be reproduced either by using a non stationary object description or through a regionalized distribution of the objects. In this paper, we focus on the latter approach.

The main contribution of the proposed method is the fact that the fit of the intensity point distribution is obtained globally in one computation step for any non stationary facies distribution. The interest is to constrain, a priori, the lithofacies simulation by a given 3D proportion distribution and not by convergence during the simulation process.

1 Introduction

Object-based models are generally used to reproduce the sedimentary units composed of geological bodies showing characteristic shapes such as channels, and for which the connectivity has to be honored. These models come up against several working difficulties. First, the definition of the objects in terms of probabilities, done by mimicking the current fluvial deposits, is inferred only from incomplete observations: 2D for horizontal plane or vertical sections from analogues, or 1D from well logs. Second, the distribution of these objects depends on the solution of the first. Indeed, the Boolean parameters are only accessible indirectly because the resulting image depends both on the objects characteristics and on their distribution: the same lithofacies proportion can be obtained using different objects with adapted distribution intensity. However these proportions are considered as classical information for sedimentary unit simulations.

As a general rule, the lithofacies proportions are variable in 3D. Their distributions have been displayed as a matrix of vertical proportion curves (VPC) (Beucher, 1992). These proportions contain the variability of the lithofacies distribution on a large scale. In the classical approach of the truncated gaussian these proportions represent the mean

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 569–578.

quantities of the lithofacies in the studied area. They are obtained by computation from well data and, if available, additional information either qualitative (knowledge of geologically homogeneous zones) (Eschard, 2002) or quantitative (in case of correlation between some lithofacies proportions and seismic-derived attributes) (Moulière, 1996). Since these proportions integrate a large part of the information needed for the construction of a geological image, it is therefore logical to evaluate the boolean parameters from this 3D proportion matrix.

2 Boolean Parameters

The Boolean model is the random set obtained by the union of objects (or objects *A*) with a given distribution, independent and located according to a Poisson point process $\wp$:

$$X = \bigcup_{x \in \wp} A(x) \; ; \qquad x \in \mathbb{R}^3$$

The model is entirely defined by the knowledge of two sets of parameters: the Poisson point process parameters and the object definition.

In this framework, local proportions can be accounted for by making object parameters variable in space, or by using a regionalized distribution of objects of the same family. In this paper, all the objects belong to the same family, their characteristics do not depend on their position; the non stationarity has been reproduced by a non stationary distribution which corresponds to a Poisson point process whose intensity *f(x)* depends on location *x*. This function is positive and the mean number of points belonging in a domain *B* is equal to $\int_B f(x)dx$.

Two examples of realizations of point processes are shown on figure 1, the first one with constant intensity, the second one with regionalized intensity.

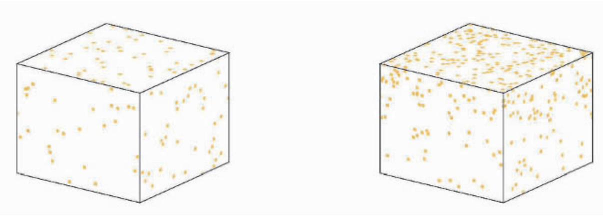


***Figure 1 -** Realizations of Poisson point process – a) using constant intensity b) using decreasing intensity from top to bottom.*

Under these conditions, the distribution of the objects is characterized by its Choquet capacity, which is the probability for any compact set *K* to intersect a set *S*:

$$T_S(K) = P(S \cap K \neq \varnothing)$$

The set *K* can be chosen as complex as wanted, but for practical reasons, in particular as available data are reduced to lines, and for fitting purposes, we only work with a set *K* reduced to a point. For a set composed of a pair of points, the computed probability corresponds to the covariance of the boolean model. In that case the fit of the experimental indicator variograms gives another relationship between *A* and *f(x*). But this relationship is only useful in a stationary case (Benito, 2003). The computation of this probability on multipoint sets can be performed when a complete image is available in Schmitt (1997).

When *K* is reduced to a point, the Choquet capacity gives particular probabilities.

When *S* = *A(x)*, we obtain $p_0(y)$, the probability at any point y, associated to an object located at x, named "object-probability":

$$T_{A(x)}(\{y\}) = P(A(x) \cap \{y\} \neq \varnothing) = P(y \in A(x)) = p_0(y)$$

When *S* = *X*, we obtain *p(y)*, the probability associated to the boolean set, named "boolean set probability":

$$T_X(\{y\}) = P(X \cap \{y\} \neq \varnothing) = p(y)$$

3 Fitting the Poisson point intensity

The fitting process consists in estimating the Boolean intensity *f(x)* knowing the "object probability" p_0 and the "boolean set probability" *p*.

The "boolean set probability" can be considered as the convolution of the intensity f by the object-probability p_0 (Schmitt, 1997 and Benito, 2002):

$$\begin{aligned} p(y) = T_X(\{y\}) &= 1 - \exp(-\int_{R^3} f(u) T_{A(u)}(\{y\}) du) \\ &= 1 - \exp(-\int_{R^3} f(u) P(A(u) \cap \{y\} \neq \varnothing) du) \\ &= 1 - \exp(-\int_{R^3} f(u) P((u-y) \in A) du) \\ &= 1 - \exp[-(f * p_0)(y)] \end{aligned}$$

The regionalized intensity *s(x)* can then be directly obtained by taking the Fourier transform (FT) of the previous equation and inverting the resulting expression:

$$s(x) = (f * p_0)(x) = -\ln(1 - p(x)) \xrightarrow{FT} \hat{s}(\nu) = \hat{f}(\nu)\hat{p}_0(\nu)$$

The difficulty is that the FT of $p_0(y)$ may attain very small values at some frequencies, making the intensity diverge. A very simple way to stabilize this operation is by means of the *Wiener filter*, a classical method used in digital image processing (González, 1992 and Pratt, 1978). Supposing the convolution affected by a white noise not correlated with the intensity, this filter gives the best estimator of the intensity function in the sense of the minimum mean square error. In this approach, the intensity and the noise are considered to be random functions of known means and spectral densities. For further computations it has been shown that the corresponding term can be taken as a constant *R*. Thus the estimation of the function f can be written as:

$$\hat{\hat{f}}(v) = \hat{s}(v)\frac{\hat{p}_0^{\,*}(v)}{|\,\hat{p}_0(v)\,|^2 + R} \qquad (1)$$

By this way, one intensity function is obtained for each couple (object-input proportions).

4 Input probabilities

The two probabilities p and p_0 involved in the fitting process are determined from the data set. In the computations, they are digitized on cells whose size is the same as that of the simulation grid.

4.1 Object probability

By definition it is the probability that a point near the germination point (or object origin) is covered by an object of the chosen family. This probability depends on the object shape and on the distribution laws of its parameters.

- From a practical point of view, the object shape is chosen according to geological qualitative information. The choice of the object depends on the working scale and on the available information. As this type of model is used for simulating fields composed of a relatively large number of objects with similar characteristics, a mean shape should be defined. For fields where very few channels are available, and also depending on the working scale, other simulation models would be more appropriate (Lopez, 2001). Depending on the lithofacies to be simulated as objects, very simple objects can be sufficient, parallelograms or ellipsoids for instance. But any kind of shapes can be used, if their different description parameters can be fitted from the available dataset; however the more complex the shape, the longer the conditional simulation.

- Knowing the different description parameters, the object probability can be evaluated. For an object of constant size, the object probability is equal to 1 near the germination point (or origin of the object) in the neighborhood corresponding to the object size, and 0 elsewhere. Figure 2 presents the 3D probability of an object of irregular shape on which the 3 main dimensions are chosen according to uniform laws.

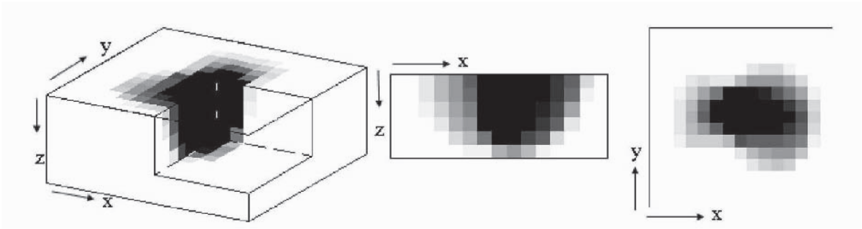


Figure 2 - *Example of object probability (from 0 in white to 1 in black) - 3D view, vertical and horizontal sections. The object origin is located at the center of the top surface.*

4.2 Boolean set probability

By definition it is the probability that a point in a domain belongs to the simulated lithofacies. This probability is indirectly known by its mean within a given volume. As stated previously, the lithofacies distribution is in fact given by the proportions that are the means of the lithofacies indicators computed on given volumes. We are not actually looking for a mean proportion on several realizations corresponding to the "point" indicator probability but for a spatial mean on each realization involving a support notion when computing the initial data and also when analyzing the results. In fact the proportions are obtained in a grid whose cell sizes are larger than those of the simulation grid. An example of the different terms is presented on figure 3. Given a very smooth 3D proportion model (figure 3a) and ellipsoidal objects, several simulations can be generated. The resulting proportions (figures 3-b) are computed for each simulation on a large grid (one proportion cell for 100 simulation cells).They globally present the same behavior as the model with a lot of variations. The same observations can be made on the probability map computed for all the simulations on the simulation grid (figure 3-c).

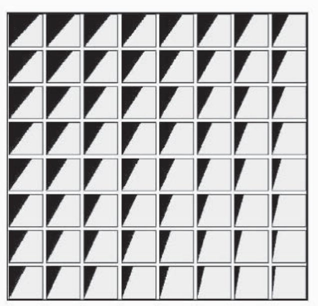


***Figure 3a** - Input proportions in the model*

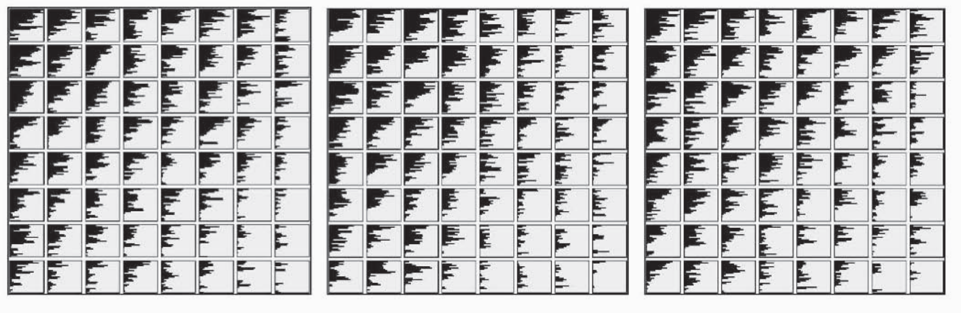


***Figure 3b** - Proportions obtained from 3 different realizations*

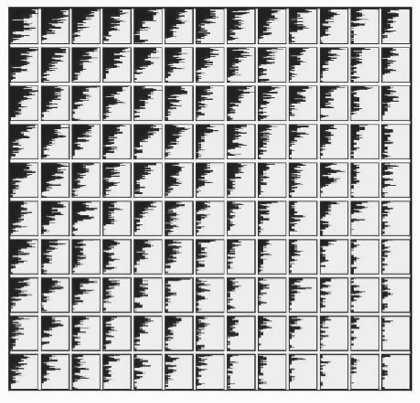


Figure 3c **-** *Probability estimated for 10 different realizations on the simulation grid*

As a first approximation and based on these previous experimental results, the probability in each cell of the simulation grid, has been taken equal to the mean known in the corresponding proportion cell.

5 From 3D proportion to 3D intensity

The 3D intensity computation is performed on the simulation grid. Knowing the object family and therefore its probability map (p_0), the estimation of signal f using formula (1) is performed and gives a 3D grid of values.
In the model, this parameter corresponds to the Poisson point process intensity which is a positive or zero number equivalent to the number of points falling into a given neighborhood. As the computation is performed without any constraint on the positiveness of the different terms, the resulting values can be negative in particular in areas where the simulated lithofacies is absent. The smoothing process chosen for correcting these values is simple to use and preserves the global proportion. However the levels where the lithofacies disappears are not clearly delimited; smoothing gives positive values on the borders and thus after the simulations are completed the resulting proportions can be positive.
Thus, before using the 3D proportion it is important to specify the meaning of the proportion data and their degree of confidence. Indeed if a null input proportion has to be strictly honored, this means indirectly that the data at these points are deterministic. Thus, before running a simulation, particular constraints such as knowledge of borders, have to be specified first.

6 Some applications

All the following simulations are conditioned by well data (Lantuéjoul, 2001) but we only focus on the non stationary distribution aspect.

6.1 Simulations of shaly lenses

In this example, the unit is composed of two main lithofacies: a reservoir lithofacies and a barrier lithofacies that modifies the fluid flow. The barrier lithofacies is composed of shaly lenses that can be simulated as ellipsoidal objects. The 3D distribution of this lithofacies is presented on figure 4, first using a 3D block image then through 4 vertical proportion curves. On this image we can see the non stationarity, in particular the depth variation of the shale levels.
Assuming that the lenses are half ellipsoidal objects, the 3D intensities are computed and presented as VPC (figure 5). The intensity distribution is of course similar to the proportion distribution.

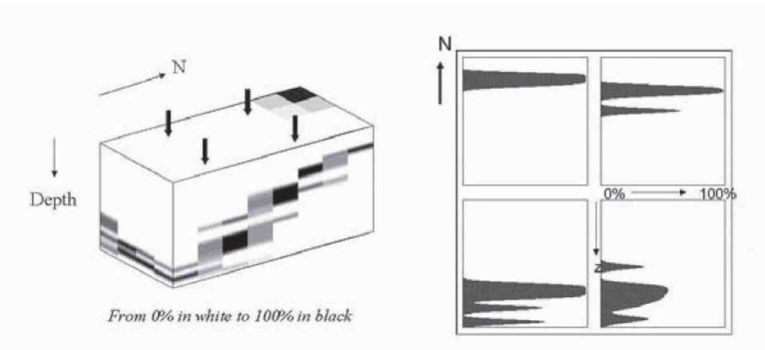


Figure 4 - *3D Shale proportions - a) in a 3D block - b) in 4 vertical proportion curves.*

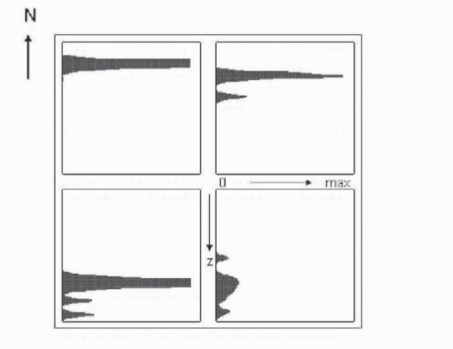


Figure 5 –*3D intensity obtained with objects of ellipsoidal shape.*

Using the mean number of objects obtained from the previous computation, simulations can be performed. One simulation is presented on figure 6.

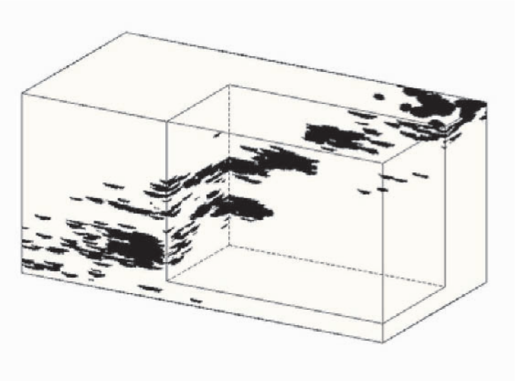


Figure 6 *– Simulation of shaly lenses*

The proportions computed on different simulations (figure 7) present, as expected, fluctuations around the distribution of the input data.

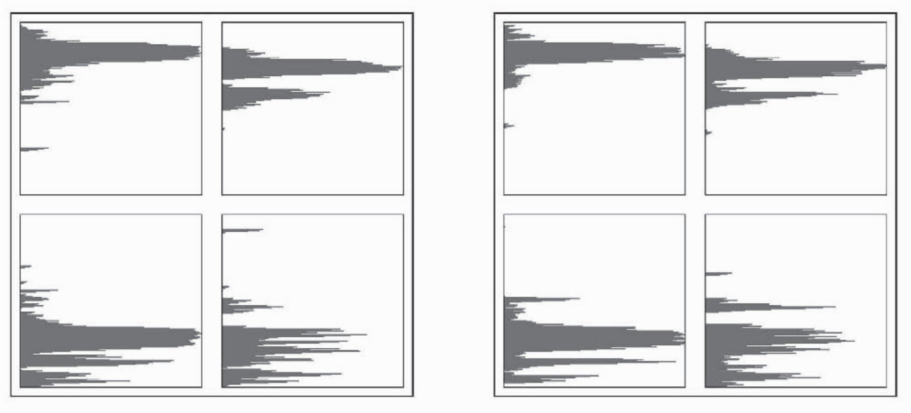


Figure *7* ***–*** *Proportions computed on 2 different simulations.*

6.2 Simulations of channels

In this second example, the problem is to simulate channels that are objects crossing the whole field at the studied scale. Thus it is impossible to give an object length. Using the previous approach, we only consider a 2D vertical section, more precisely a plane more or less perpendicular to the flow. On such a section the channels are limited bodies whose non stationary distribution can be evaluated as previously described. The proposed approach consists in fitting the Boolean parameters on a representative vertical section and then using the third dimension parameters (orientation, horizontal shape) to propagate the channel from this section to the whole volume. This method can only be applied to deposits where channels run practically parallel to each other, that is when the number of channels is more or less constant in the whole working area.

For illustration purposes, a synthetic dataset has been created. In the 2D vertical section, the channel facies (in black) is located at two distinct levels (figure 8) and its proportion decreases from west to east. A simulation has been performed after computing the 2D intensity (figure 9). The result is satisfactory even though there is room for improvement in fitting the horizontal parameters.

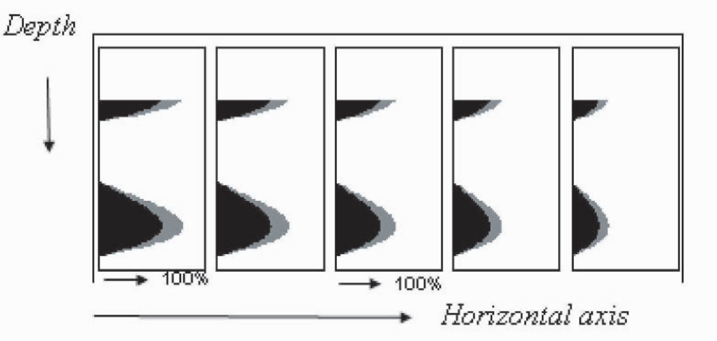


Figure 8 - *Vertical proportion curves along a cross section perpendicular to the flow*

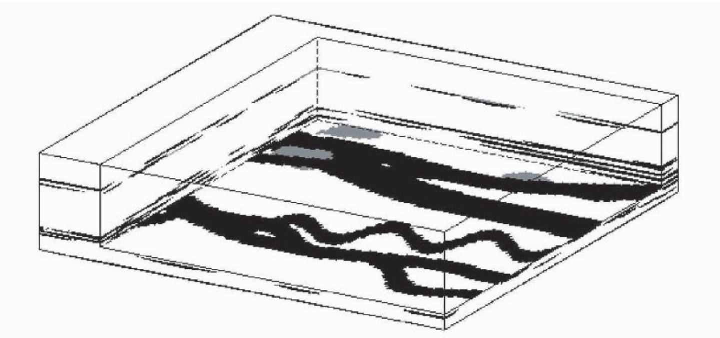


Figure 9 - *Simulation of channels conditioned by the previous vertical 2D non stationarity proportion curves*

7 Conclusions and perspectives

Given the object distribution, the approach presented here to estimate the Poisson intensity from 3D proportions gives satisfactory results although several improvements and extensions can be considered.

- For the time being, the proportions are predicted by kriging. Their simulation would be more appropriate.
- Knowledge of the object distribution being difficult to attain, a realistic alternative is to replace the object distribution by a family of distributions, which suggests a Bayesian framework.

References

BENITO GARCÍA-MORALES M., BEUCHER H. 2002. Inference of the boolean model on a non stationary case. In *Proceedings of the IAMG Annual Conference, Berlin'02*, vol. 1, p. 167-172.

BENITO GARCÍA-MORALES M. 2003. Non stationnarité dans les modèles de type boolèen : Application à la simulation d'unités sédimentaires. PhD in geostatistics Ecole des Mines de Paris.

BEUCHER H., GALLI A., LE LOC'H G., RAVENNE C. & Heresim Group. 1993. Including a regional trend in reservoir modelling using the truncated Gaussian method. In A. Soares (ed.) *Geostatistics Tróia '92*, Kluwer, 1993, vol. 1, p. 555-566.

ESCHARD R., DOLIGEZ B., BEUCHER H. 2002. Using quantitative outcrop databases as a guide for geological reservoir modelling. In M. Armstrong et al. (eds.), *Geostatistics Rio 2000,* Kluwer, 2002, p. 7-17

GONZÁLEZ R.C., RICHARD E. 1992. *Digital Image Processing*, Addison-Wesley.

LANTUÉJOUL C. 2001. *Geostatistical simulation : models and algorithms*. Springer.

LOPEZ S., GALLI A., COJAN I. 2001. Fluvial meandering channelized reservoirs : a stochastic and process-based approach. *In :* 2001 Annual Conference of the International Association for Mathematical Geology, Cancún (Mexico), September 6-12, 2001. [CD-ROM]. s.l.n.d. 16 p.

MOULIÈRE D, BEUCHER H., HU L.Y., FOURNIER F., TERDICH P., MELCHIORI F. and GRIFFI G. 1997. Integration of seismic derived information in reservoir stochastic modelling using truncated Gaussian approach. In E.Y. Baafi and N.A. Schofield (eds.), *Geostatistics Wollongong '96*, Kluwer , 1997, vol. 1, p. 374-385.

PRATT W.K. 1978. *Digital Image Processing*. Wiley.

SCHMITT M., BEUCHER H. 1997. On the inference of the Boolean model. In E.Y. Baafi and N.A. Schofield (eds.), *Geostatistics Wollongong '96*, Kluwer, 1997, vol. 1, p. 200-210.

FINE SCALE ROCK PROPERTIES: TOWARDS THE SPATIAL MODELING OF REGIONALIZED PROBABILITY DISTRIBUTION FUNCTIONS

MICHEL GARCIA[1], DENIS ALLARD[2], DAVID FOULON[3], and SYLVIE DELISLE[4]
(1) *FSS International r&d, 1956 avenue R. Salengro, 92370 Chaville, France*
(2) *INRA, Unité de Biométrie, Domaine St-Paul, 84914 Avigon cedex 9, France*
(3) *TOTAL, Tour Coupole, La Défense 6, 92078 Paris La Défense, France*
(4) *TOTAL, CSTJF, avenue Larribau, 64018 Pau Cedex, France*

Abstract. Conventional logs and borehole images are common sources of short-scale property data for the characterization of petroleum reservoirs. Short-scale properties are consequential for production performance purposes but are also representative of too small rock volumes for direct full-field modeling. Average, possibly facies-based, properties are generally derived and stochastic simulation methods are used to simulate them throughout the whole reservoir. The regionalization of short-scale property distribution functions is considered here to take into account the short-scale variability of data. The novelty of the proposed approach is to regionalize mixtures of distributions and to correlate them to secondary information. Application to simulate distributions of fracture-frequencies in a naturally fractured reservoir illustrates the approach.

1 Introduction

Conventional logs and borehole images on vertical, deviated or horizontal wells are common sources of information in petroleum reservoir characterization. They provide a short-scale description of rock properties, content or discontinuities (e.g. fractures). The short-scale description resides in the small support of data (i.e. the representative rock volume) associated with a small sampling spacing along wells. Typically, support and sampling spacing do not exceed tens to hundreds of centimeters for data that are recorded over tens to thousands meters long well intervals. This results in numerous and valuable data about reservoir heterogeneities. The integration of such data into a full-field reservoir model raises, however, a problem of support effect. High-resolution models, with grid-cell volumes as small as the representative volume of data, mean a dissuasive high number of cells, beyond the capacity of present-day computers. The usefulness of such very fine grids is also questionable.

The problem of different scales of heterogeneity and data in reservoir characterization has already been the focus of attention of several authors (see for example Tran, 1995, for a literature review). It is unanimously recognized that the short-scale heterogeneity strongly affects the macro-scale flows and particularly multiphase flows. Nonetheless, short-scale properties are generally averaged on wells to be used as conditioning data to

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 579–589.

model scalar variables. The latter can be interpreted either as a cell property or as a point property related to a larger support. For these scalar variables, spatial statistics are inferred and conventional geostatistical methods are used to estimate or simulate them within the reservoir. A local uncertainty can then be derived at any location as a local distribution of cell-related (average) values. This local distribution is of course different from the one of short-scale measurements that would have been obtained if a well had been drilled and logged at this location. It follows that beyond the averaging or calibration step, the short-scale variability of data is not exploited nor modeled.

There are situations, however, that would justify considering as regionalized variable the whole distribution of short-scale values instead of a mere (average or calibrated) variable. Typically, when independent or poorly correlated phenomena (or genetic processes) can explain the short-scale variability of property values, the distribution of values can be seen as a mixture of distribution components (populations) each related to one phenomenon. The availability of structural, seismic, or other type of exhaustive information, which could be related to at least some of these phenomena, then makes it possible to use them as secondary information to model parameters or features of the corresponding distribution components. Possible situations are as follows.

- In naturally fractured reservoir, a same strike-direction of fractures may correspond to diffuse fracturing or fracturing swarms (corridors). The two types of fracturings are related, however, to different geological or tectonic episodes.
- Whether rock properties are facies-dependent but the facies-types are badly defined, the variability of within-facies properties is important, or simply as an alternative to prior modeling of facies-types or of proportions of facies, the distributions of short-scale properties can be directly modeled as regionalized variables.

Although the spatial modeling of distribution functions or more generally of non-scalar variables is poorly exploited in reservoir characterization, it has been tackled already in other (environmental) domains to model for instance soil properties or time-series in air pollution. This paper revisits this avenue in the context of reservoir characterization as a way to exploit better short-scale logs or borehole imaging data. Section 2 poses the problem and reviews the possible approaches to make a distribution function a random variable. Section 3 presents the proposed approach in which distributions of short-scale variables are identified to mixtures of distributions. Section 4 addresses aspects of model fitting for a mixture of distributions. A naturally fractured reservoir case study illustrates the approach in Section 5 and conclusions are drawn in Section 6.

2 Spatial modeling of distribution functions

A geostatistical approach is sought to model regionalized distribution functions of short-scale rock properties. It must be adapted to the following conditions and objectives.

- Well logs or borehole images are sources of short-scale property data from which sample distribution (histograms) can be calculated over wells or portions of wells.
- The support of the distribution is given by the selected well-portion length used to calculate the distributions. It must be consistent with grid-cell dimensions.
- Spatial changes of the shape, spread or modes of the distributions can be explained

by different phenomena related to available secondary information.

- The expected results are the estimation or simulation of distribution functions at non-sampled locations. Quantification of the uncertainty about these distributions is also desired. This calls for a notion of distribution of distributions.

Point or block estimates are commonly used in geostatistics. They involve scalar variables which are associated with random functions (regionalized variables). Looking at probability distribution functions, “functional” or vectorial variables are to be spatially modeled. Though some avenues have been proposed in the literature to regionalize directly functional variables (see for example Goulard and Voltz, 1992), only vectorial variables are considered here. The vectorial variable may be a set of parameters (continuously defined distribution function) or a set of prescribed values for which the distribution function is calculated (discretely defined distribution function). In both cases, the components of the vectorial variable are to be coregionalized in space.

The choice of an approach is closely related to the way the distribution function is defined. Parametric, semi-parametric, non-parametric or mixture-based approaches can be considered. They are briefly discussed hereafter.

2.1 PARAMETRIC OR SEMI-PARAMETRIC APPROACH

The basis for a parametric or semi-parametric approach is the choice of a general enough function that can span a variety of distribution shape, modes, spread and location with a minimum number of parameters. “Parametric” is generally used to refer to an existing model of distribution function (e.g. a beta law), “semi-parametric” applying to any other type of function with good properties (e.g. polynomial function). The function is randomized by making its parameters a vector of random functions. Parameters of the function are to be calculated first to fit the sample distributions at well locations. Maximum-likelihood (for parametric functions) or least-squares techniques are generally employed at this stage. Traditional geostatistical methods can then be used to estimate or simulate (jointly) the parameters elsewhere. This type of approach was already proposed by some other authors. We can mention Goulard and Voltz (1992), who use polynomial spline functions to model soil water retention curves, or Kyriakidis and Journel (1998) who apply deterministic temporal trend functions to sulfate deposition data from monitoring stations. If the simplicity of such an approach makes it attractive, parametric functions remain limited in shapes and semi-parametric functions raise a number of potential difficulties: (1) least-square fitting of a semi-parametric function requires that the sample distribution be entirely defined (no partial data), (2) interpretation of the parameters to relate them to distribution populations and secondary variables may not be straightforward, (3) complexity of the interpretation is even higher if the parameters are correlated or are to be transformed to make them uncorrelated.

2.2 NON-PARAMETRIC APPROACH

Non-parametric approaches call for a discrete form of the distributions. Based on a “curve sampling” framework, the distribution functions are calculated only for a given number of selected short-scale property values. The so defined discrete distribution functions are then regionalized through the vector of distribution function values (i.e. probabilities) that are made random functions (RF).

A few references can been found in the literature on non-parametric geostatistical approaches to modeling probability distributions or curves. Goulard and Voltz (1992) tried a non-parametric approach to model the soil water-retention curves previously mentioned. Cokriging is used to estimate fields of correlated discrete curve values (= RF vector) from which the entire curve can be rebuilt at each estimation point. Uncertainty about the so estimated curves is not addressed. In a different way, Desbarats (2000) proposes an approach to simulate 3D spatial fields of sediment grain-size distributions. The weight percentages of different grain-size classes comprise the vector of regionalized variables. The minimum/maximum autocorrelation factors (MAF) method is used to transform the initial RF vector into a vector of spatially uncorrelated (MAF) RFs. Sequential Gaussian simulation is used to simulate independently the MAF RFs.

The main advantage of non-parametric approaches is their ability to reproduce any distribution shape. They can also integrate partial data, i.e. not fully defined sample distributions. The main difficulty is the existence of (possibly complex) spatial cross-correlations between the RFs of the probability vector, i.e. between the discrete values of the regionalized function or curve. The integration of secondary information is also another concern. It introduces additional variables to be correlated to the already cross-correlated (probability) RFs. Looking at "de-correlation" solutions like the MAF method, other problems do rise: (1) general efficiency of the method, (2) amount of data required to obtain reliable (multivariate-processing) results, (3) interpretation of the uncorrelated RF transforms against the available secondary variables.

2.3 APPROACH BASED ON A MIXTURE OF DISTRIBUTIONS

Mixture of distributions here refers to a distribution function written as the sum of two or more parametric distribution functions. Each (elementary) parametric distribution is also called a "distribution component" of the mixture and is associated with a "mixing proportion" which determines the contribution (or relative frequency) of this component in the overall distribution. With this definition, a mixture of distributions can be seen as a semi-parametric approach at the important difference that each distribution component is clearly identified and can be characterized, if necessarily, independently of others (accounting for incomplete data). In addition, all parameters may have a statistical meaning. It follows that estimation theory methods can be used to calculate the parameters of a distribution mixture from a sample distribution (e.g., maximum-likelihood estimation). These parameters can be readily interpreted and related to explicative secondary variables. This is the type of approach that was finally retained to model distributions of short-scale properties. Details about the approach and application to simulate distributions of fracture frequencies are the purpose of the following sections.

3 Regionalization of a mixture of distributions to model short-scale properties

A mixture-based approach is sought to model the spatial uncertainty about the distributionofashort-scale property *Z*. Properties and distributions are related to different measurement and observation scales, respectively. Let denote *v* the short measurement scale along wells and V the coarser scale at which property distributions are evaluated. If wells are vertical, the volume V is representative of vertical property profiles. If wells are deviated or horizontal and the distributions are calculated over similar portions of wells, V is to be interpreted more as a neighborhood around each well portion location. Mathematically, the distribution function φ of the short-scale property Z, measured on a support v within a neighborhood V centered at location $\mathbf{u}$, can be written:

$$\varphi(z;\mathbf{u}) = \varphi_{Z(v)}(z;V(\mathbf{u})) .$$

where φ denotes indifferently the probability density function (pdf) or the cumulative distribution function (cdf) of Z within the local neighborhood $V(\mathbf{u})$. Its decomposition (or approximation) into a mixture of N_c distribution components can be expressed as:

$$\varphi(z;\mathbf{u}) = \sum_{k=1}^{N_c} p_k(\mathbf{u})\varphi_k(z;\boldsymbol{\theta}_k(\mathbf{u}))$$

where $p_k(\mathbf{u})$ is the mixing proportion at location $\mathbf{u}$ of the k^{th} parametric distribution component φ_k. The mixing proportions must range from 0 to 1 and sum to 1. The choice of a parametric distribution model for each component determines the parameter vector $\boldsymbol{\theta}_k(\mathbf{u}) = (\theta_{k,1}(\mathbf{u}),\ldots,\theta_{k,n_k}(\mathbf{u}))^T$. For example, a Gaussian component φ_k will be associated with a mean and a variance, i.e. $\boldsymbol{\theta}_k = (\theta_{k,1} = \mu_k, \theta_{k,2} = \sigma_k^2)^T$. Regionalization of the mixture of distributions comes to make the mixing proportions and the distribution-component parameters a set (vector) of RFs, namely:

$$\text{regionalized } \varphi(z;\mathbf{u}) = \Phi(z;\mathbf{u}) = \sum_{k=1}^{N_c} P_k(\mathbf{u})\varphi_k(z;\boldsymbol{\Theta}_k(\mathbf{u}))$$

where $P_k(\mathbf{u})$ and $\boldsymbol{\Theta}_k(\mathbf{u}) = (\Theta_{k,1}(\mathbf{u}),\ldots,\Theta_{k,n_k}(\mathbf{u}))^T$ are all RFs. These RFs are to be estimated or simulated at the nodes of a (regular) grid to honor:

- data about the mixture parameters $p_k(\mathbf{u}_i)$ and $\varphi_k(\mathbf{u}_i)$ at well locations $\mathbf{u}_i$,
- correlations between mixture parameters or with external secondary variables.

For consistency purposes, the grid-node spacings are expected not to be smaller than the well portion lengths used to calculate the sample distributions. Regarding well data, not all mixture parameters $p_k(\mathbf{u}_i)$ and $\varphi_k(\mathbf{u}_i)$ are to be known at each location. For example, on a “short” well, relevant statistics could be available only for one or a few distribution components corresponding to the populations crossed by the well.

Estimated or simulated fields of $P_k(\cdot)$ and $\Theta_k(\cdot)$ are intended to rebuild fields $\Phi(z;\cdot)$ of the distribution function of the short-scale property Z. Quantifying the spatial uncertainty about $\Phi(z;\cdot)$ is also another issue. These objectives are more easily attained by using a stochastic simulation approach. Realizations of $\Phi(z;\cdot)$ are obtained by just combining jointly or independently simulated realizations of the mixture parameters. The steps of the approach can be summarized as follows.

1. Choice of a number and models of distribution components.
2. Estimation of mixture parameters to fit sample distribution from well portions.
3. Inference of statistical models for each mixture parameter (RF).
4. Joint or independent stochastic simulation of the mixture parameters.
5. Post-processing of the simulated fields of mixture parameters.

In step 1, the distribution components are to be related to verified or guessed genetic processes or phenomena that explain short-scale property populations. In step 2, robust methods are required to estimate mixture parameters from sample distributions. A maximum-likelihood-based method is presented in section §4.1. Instead of processing sample distributions, an alternative consists in analyzing logs data to recognize significant modes, i.e. successive increases and decreases of $Z(v)$ that can be related, with confidence, to different populations along wells. Such an approach was developed and successfully applied to identify fracture corridors from fracture-frequency logs. It is based on the SiZer (SIgnificant ZERo crossings of derivatives) approach introduced by Chaudhuri and Marron (1999) to test the presence of modes in a pdf.

Steps 3 and 4 involve more traditional geostatistical tasks. Especially, any stochastic simulation method, whether based on two or multiple-point statistics but able to integrate secondary information, can be used to simulate any mixture parameter. The only particularity here is that the statistical models in step 3 are to be established from estimated instead of measured data. More precisely, the measured data are short-scale properties from which mixture parameters (i.e. statistics) are to be calculated over well portions. All well portions do not necessarily have the same orientation nor the same number of short-scale data, hence the estimates may be imprecise and the precision different. Estimation methods of sample histogram or variogram should ideally take into account the precision of the data. Preliminary results have been obtained for the estimation of the variogram of the mean of a distribution component. They are presented in §4.2. Regarding step 5, a realization of $\Phi(z;\cdot)$ is obtained from a realization of each of the mixture parameter. Cross-correlated parameters are, however, to be jointly simulated and realizations from the same simulation are to be taken together. For independent parameters or groups of parameters, any simulation of one group can be combined with a realization of another group. A multivariate Latin hyper-cube (Monte-Carlo) sampling technique can then be used to span more efficiently the space of variability of $\Phi(z;\cdot)$.

4 Aspects of statistical model fitting for a mixture of distributions

4.1 EM ALGORITHM FOR ESTIMATING MIXTURE PARAMETERS

The Expectation-Maximization (EM) algorithm is a general iterative algorithm that computes maximum likelihood estimates (MLEs) of the parameters of a mixture of distribution components when the group (distribution component) memberships of the data are unknown (missing information). A detailed presentation can be found in the seminal paper by Dempster et al. (1977). McLachlan and Krishnan (1997) provide a review with examples and extensions.

The general idea of the EM algorithm is to compute iteratively the augmented likelihood of the data, namely, that taking into account the missing information. Starting with some initial values, the expectation step (E-step) computes the conditional expectations of the augmented likelihood, i.e. the conditional probabilities for the data to belong to the different groups (distribution components) given the current values of the mixture parameters. The maximization step (M-step) computes the MLEs of the mixture parameters, given the measured data and the updated conditional probabilities from the E-step. Dempster et al. (1977) showed that each step increases the augmented likelihood. These two steps are iterated until convergence occurs at a local maximum of the likelihood surface.

Clustering methods of independent data based on mixtures of normal (Gaussian) distributions coupled with an EM algorithm have been shown to be powerful, see for example McLachlan and Basford (1988) and Banfield and Raftery (1993). The data are supposed to originate from a pdf mixture $f(z)=\sum_{k=1}^{N_c} p_k f_k(z;\boldsymbol{\theta}_k)$ where $\boldsymbol{\theta}_k$ is the parameter vector of the k$^{\text{th}}$ distribution component and p_k the mixing proportion. Independent data being assumed, it can be shown that the E-step at iteration (q) is equivalent to estimating the conditional probabilities:

$$\hat{t}_{ik}^{(q)} = P(z_i \in \text{group } k \,\big|\, (\hat{\boldsymbol{\theta}}_1^{(\mathrm{q})},...,\hat{\boldsymbol{\theta}}_{N_c}^{(\mathrm{q})})) = \frac{\hat{p}_k^{(q)} f_k\left(z_i;\hat{\boldsymbol{\theta}}_k^{(q)}\right)}{\sum_{l=1}^{N_c} \hat{p}_l^{(q)} f_l\left(z_i;\hat{\boldsymbol{\theta}}_l^{(q)}\right)} \tag{1}$$

where $i = 1,\ldots, n$ (number of data) and $k = 1,\ldots, N_c$. After an E-step, the classification matrix is not a 0/1 matrix but each row of the matrix still adds up to one. For Gaussian variables, the maximum likelihood estimators are equivalent to the method of moment estimators. Hence, the M-step at iteration (q + 1) gives:

$$\left.\begin{array}{l} \hat{n}_k^{(q+1)} = \sum_{i=1}^{n} t_{ik}^{(q)}, \quad \hat{p}_k^{(q+1)} = \hat{n}_k^{(q+1)} \big/ n \\ \hat{\mu}_k^{(q+1)} = \sum_{i=1}^{n} t_{ik}^{(q)}\, z_i \big/ \hat{n}_k^{(q+1)}, \quad \left(\hat{\sigma}_k^2\right)^{(q+1)} = \sum_{i=1}^{n} t_{ik}^{(q)}\, z_i^2 \big/ \hat{n}_k^{(q+1)} - \left(\hat{\mu}_k^{(q+1)}\right)^2 \end{array}\right\} \tag{2}$$

For fracture frequencies related to Gamma variables, the equivalence between maximum likelihood and method of moment is only valid for the first moment, i.e. the α parameter of the gamma distribution when $\beta = 1$. We can then use the first two equations of (2) for estimating α. If $\beta \neq 1$, we made the approximation that the variance can be also estimated using (2). This leads to the estimators $\hat{\alpha} = \hat{\mu}^2 \big/ \hat{\sigma}^2$ and $\hat{\beta} = \hat{\sigma}^2 \big/ \hat{\mu}$. In all tested situations, convergence holds and the parameters are correctly estimated.

4.2 ESTIMATION OF THE VARIOGRAM OF THE RANDOM FIELD $\mu(\cdot)$

We consider the problem of estimating the variogram of the random field $\mu(\cdot)$ whose variogram is $\gamma(h)$ and mean is m. It is recalled that in each neighborhood (or block) $V(\mathbf{u})$ independent data $Z(v_i)$ are assumed (they correspond to data measured on a smaller scale v). The problem is that this field is never sampled directly; we only have estimations $\hat{\mu}(V)$ of this field in neighborhoods $V(\mathbf{u})$ at different locations. These estimations are unbiased but they have different degrees of precision, depending on the number of available data in the neighborhood $V(\mathbf{u})$ (or more precisely along a well portion) for estimating $\hat{\mu}(V)$. We consider the case where there are n_k data in neighborhood V_k = V($\mathbf{u}_k$) for estimating $\hat{\mu}(V_k)$. The sample variogram is $\hat{\gamma}(h) = \frac{1}{2Nh}\sum_{N_h}\{\hat{\mu}(V_k) - \hat{\mu}(V_l)\}^2$.

Proposition 1: *For the model above,* $\frac{1}{2}E\left[\{\hat{\mu}(V_k) - \hat{\mu}(V_l)\}^2\right] = \frac{n_k + n_l}{2n_k n_l}m + \gamma(h)$.

Proof: Available from the authors.

5 Application to a naturally fractured reservoir

A mixture-based approach was applied to model distribution functions of fracture frequencies within a naturally fractured oil reservoir. The reservoir is a 10 km wide pop-up shaped structure and internal faults occur with directions parallel and transverse to the fold axis. Its tectonic history explains several fracture orientations. Fractures exist throughout the whole 1000 m thick upper part of the reservoir, in all lithofacies but possibly with different occurrences. The analysis of interpreted fractures from core plugs, cores and borehole images reveals three main fracturing levels.

- Micro-fractures and fissures that connect matrix vugs and provide a background (matrix) permeability (k) of about 0.01 mD.
- Diffuse fractures that provide the main connectivity in the reservoir (k ≈ 10 mD).
- Fracture swarms, interpreted as parallel fractures vertically extended over tens of meters. They provide the most productive intervals (k > 100 mD).

The two latter types of fractures are the ones observed from borehole images. Interpreted fractures from the borehole images acquired on four wells were first classified into directional fracture-sets using a cluster analysis method. Three such fracture-sets were retained for this reservoir. Fracture frequencies were then calculated for each directional fracture-set. The fracture frequency (FF) is defined as the number of fractures per unit length measured perpendicularly to the fracturing plane. It can be seen as a *fracture density* geometrically corrected to be independent of the well directions (e.g. see Gauthier et. al., 2002). FFs are calculated using a moving window along fracture logs. The moving-window length determines the measurement scale of FF: the longer the window length, the large the support of FF.

Sets of diffuse fractures and of fracture swarms having the same orientations, a small (2 m long) moving window was used to allow distinguishing between FFs related to one or the other type of fracturing. The optimal window size was derived from a SiZer-like analysis of the fracture logs. This analysis proved very efficient to identify and locate

FF modes corresponding to fracture swarms along the wells. Based on a small moving window length, so calculated FFs can be seen as short-scale properties. The four wells were divided into nine, about 50-m-long (horizontally), well portions and sample distributions of FF were calculated for each well portion and for each fracture-set. These distributions are likely to include two populations of FFs corresponding to different diffuse and in-swarm fracturing processes. Both fracturing processes can be considered as Poisson processes but with a different rate (or intensity), higher rates being expected for fractures in swarms. It follows that the FF associated with each fracturing process should be distributed according to a Gamma distribution with shape and scale parameters $\alpha = \mu/L$ and $\beta = L$, where L = moving window length and μ = (process-specific) mean of FF. This leads to the following model of distribution mixture:

$$\varphi(z) = pG_d(z;\alpha_d,\beta = L) + (1-p)G_s(z;\alpha_s,\beta = L)$$

where the unknown parameters to be regionalized are α_d for the Gamma distribution G_d associated with the diffuse fracturing, α_s for the one associated with the fracturing in swarms, and p the (single) mixing parameter. This model was fitted to the sampled distributions using the EM algorithm to derive data about the mixture parameters.

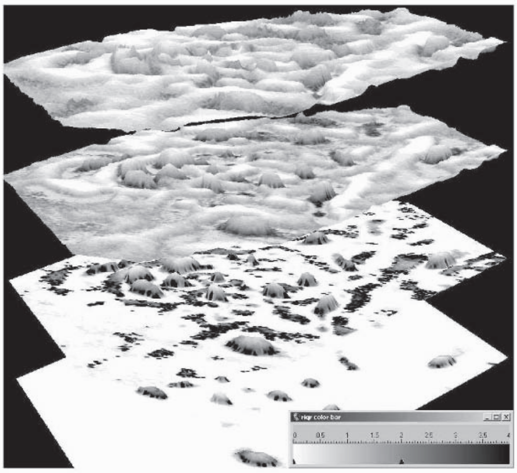


Figure 1. Quantile-surfaces with, from top to bottom, $q_{.95}$, $q_{.90}$, $q_{.75}$, and $q_{.50}$.
The elevation of the surfaces is given by the median of simulated distribution quantiles.
The gray-scale shows the relative inter-quartile range $RIQR_q = IQR_q / M_q$.

The reservoir being homogeneously fractured vertically, 2D sequential Gaussian simulation was carried out to simulate jointly the three distribution mixture parameters. For one of the fracture-set, parameter α_d was correlated with a structural curvature attribute, p with the distance to fault, and α_s with α_d. One hundred realizations were simulated for each parameter and combined to build fields of the overall distribution of FF. Quantile surfaces were derived as shown in Fig. 1. The elevation of the surfaces is given by the (quantile) median M_q and the color scale of the surfaces shows the relative inter-quartile range defined as $RIQR_q = IQR_q / M_q$. Humps and hollows on the surfaces indicate more or less likely fracture-swarm locations. Zones of high $RIQR_q$ depict zones where the uncertainty is high about the overall distribution of *FF* for the quantile considered. So obtained realizations of FF distributions can be used to constrain discrete-fracture network models and to calculate equivalent flow properties of the fracture network.

6 Conclusions

The spatial modeling of short-scale property distribution functions may be justified to improve reservoir models or as an alternative to more traditional facies-based geostatistical approaches. The proposed approach is based on distribution mixtures that are regionalized by making their parameters a vector of random functions. These parameters have a statistical meaning and can be inferred from sample distributions by using estimation theory methods. Provided the components of a distribution mixture can be related to different populations (e.g. genetic processes), the parameters can be easily interpreted and correlated to secondary variables. Partial (incomplete) distribution data can be also integrated. The approach has been applied to simulate fracture-frequency distributions in a naturally fractured reservoir. The results are simulations of distribution functions from which the local uncertainty about the distributions can be evaluated. This calls for new analysis and visualization tools.

Non fully resolved aspects of the approach concern the precision of the data that are not measured but estimated data about distribution mixture parameters. Spatial statistics are to be calculated from imprecise data and statistical models are to be inferred. Preliminary results have been established to estimate the variogram of the mean of a distribution component. Research work is still needed to investigate further the consequences of imprecise data on statistical model inference and the way models are to be corrected.

Acknowledgements

We are grateful to Dimitri D’Or (FSS International) who actively participated to the work and implemented the EM algorithm and the co-simulation program. We also thank Total SA. for its financial support and for authorizing the publication of this paper.

This page intentionally blank

References

Banfield, J.D. and Raftery, A.E. (1993) Model-based gaussian and non gaussian clustering, *Biometrics*, **49**, 803-821.

Chaudhuri, P. and Marron, J.S. (1999) SiZer for exploration of structures in curves, *J. Amer. Statist. Assoc.*, Vol. **94**, 807–823.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977) Maximum likelihood from incomplete data via EM algorithm (with discussion), *Journal of the Royal Statistical Society*, Ser. B **39**, 1–38.

Gauthier, B.D.M., Garcia, M. and Daniel, J.-M. (Aug. 2002) Integrated fractured reservoir characterization: a case study in a North Africa field, *SPE Res. Eval. and Engr.*, 284-294.

Goulard, M. and Voltz, M. (1992) Geostatistical Interpolation of Curves, in *Geostatistics Tróia '92*, Kluwer Academic Publishers, Volume **2**, 805-816.

Kyriakidis, P. C. and Journel, A. G. (1998) Stochastic Modeling of Spatiotemporal Distributions: Application to Sulphate Deposition Trends over Europe", in *geoENV II*, Geostatistics for Environmental Applications, Valencia, Spain, Kluwer Academic Publishers, 89-100.

McLachlan, G.J. and Basford, K.E. (1988) *Mixture Models: Inference and Applications to Clustering*, New York: Marcel Dekker.

McLachlan, G.J. and Krishnan, T. (1997) *The EM algorithm and Extensions*, New York: Wiley.

Tran, T. (1995) *Stochastic Simulation of Permeability Fields and their Scale-up for Flow Modeling*, Ph.D. Thesis, Stanford University

A COMBINED GEOSTATISTICAL AND SOURCE MODEL TO PREDICT SUPERPERMEABILITY FROM FLOWMETER DATA: APPLICATION TO THE GHAWAR FIELD

JOE VOELKER and JEF CAERS
Department of Petroleum Engineering, Stanford University, 367 Panama St., Stanford, CA. 94305-2220

Abstract. Discrete fracture networks are an important hydraulic component of the Arab-D Formation, in the giant Ghawar Field, Saudi Arabia. One manifestation is superpermeability, or super-k, which provides for problematic, anomalously high, localized fluid conductivity, and has eluded simple geologic characterization. Super-k spatial prediction is desired for optimal placement of water injection and oil production wells. Our super-k model, comprised primarily of discrete fracture networks, is constructed as part of a new optimization algorithm which conditions a combined fracture / facies model to both static and dynamic data. The principal component of our algorithm, the inversion of dynamic well production data to optimize super-k parameters, is based upon a new approach to the forward modeling of fracture flow: the use of sources. The source model offers a key advantage over the traditional, discretization approach, in that commercial reservoir flow simulators may be utilized for the forward simulation. Thus, the practical characterization of discrete fracture networks is achieved.

1 Introduction

The characterization of fractured rock formations, specifically their fluid conductivity properties, has application in a variety of problems, most notably of late in contamination transport in fresh water aquifers, nuclear waste storage, and in hydrocarbon production, the topic of this paper. The most general hydrocarbon production problem is the determination of optimal well placement schemes, a task requiring accurate production forecasting under a given well development plan. Although it is not uncommon for a rather homogeneous characterization of the spatial distribution of reservoir storage and conductivity to suffice for an adequate production forecast, the occurrence of either of two particular geologic phenomena often invokes the requirement for a more detailed hydrocarbon reservoir characterization. The phenomena are reservoir fracturing, and aquifer expansion. The combination of these two phenomena renders the task of rigorous characterization as inescapable. Rarely do the fracture systems in a reservoir not require consideration in generating production forecasts. Small scale, pervasive fractures,

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 591–600.

for example, may entirely define the conductivity of a reservoir. Large scale faults may be principal reservoir flow barriers, partitioning the reservoir into hydraulically isolated regions, each requiring a separate well development. The incremental development requirement has singularly cast many otherwise profitable developments as infeasible and therefore abandoned. The expansion of an associated aquifer upon hydrocarbon depletion, or equivalently, the injection of water for enhanced hydrocarbon recovery, provokes the need for a detailed characterization, simply because a fluid that generates a revenue deficit is introduced into the productive region of the reservoir. The length of time before which a conductive region may bring water to a production well, is a critical economic factor. Highly conductive regions which benefit oil and gas production, may provide detrimental pathways for water flow upon aquifer encroachment. Water flow in these pathways, accelerates costly arrival at the producing wells, and possibly circumvents the sweep of oil, or prevents the production of free gas, in significant volumes of reservoir rock. Mitigation of this problem is subsequent preferential placement of wells, in patterns designed to minimize the deleterious flow effects of these conduits.

By far the most potentially conductive elements of a reservoir are its laterally well connected discrete fracture systems, as permeability upper bounds of an extensive discrete fracture system are orders of magnitude larger than that in porous media. The characterization of these local, high conductivity geologic elements, certainly present in all rock formations to some degree, is therefore critical, albeit extremely difficult, due to geometry. Discrete fracture networks are infinitesimally thin, and occur generally as vertical planes. Therefore, direct sampling by vertical wells under conventionally large oil field areal spacing is extremely rare, although sampling in horizontal wells may be substantial. In this case, borehole imaging data, as well as core and log data, may correlate to discrete fractures. Correlation with seismic data is also difficult unless the fracture system generates significant vertical offset of formation surfaces. Seismic data is, however, critical in that the detection of faults proves the existence of a sufficient condition for the occurrence of smaller scale fracture systems. Furthermore, detailed 3D definition of the reservoir structure aids in the generation of rock fracture stress-strain models, which may prove extremely beneficial. However, the most effective characterization data for discrete fracture systems is precisely that which brings their importance to bear, production data. Given that fracture systems influence this data through a physical process, the data must necessarily be inverted to yield realizations of fracture system parameters. This is practically accomplished through an algorithm which calls iterates of simulations of the physical process, toward finding parameters that minimize the error between the dynamic data and computation.

We have attempted to characterize a fracture dominated anomaly called "super-permeability," or super-k, in the Arab-D Formation, in a small study area in the Ghawar Field, Saudi Arabia, in which the operator, Saudi Aramco, is attempting to resolve observed, massive, unmitigated hydraulic conductivity between injection and production wells separated by up to 1 km. Premature oil well abandonment has become a problem, due to the inability to predict the spatial distribution of super-k, which consequently channels injection water preferentially to adjacent production wells. Super-k prediction would enable the placement of wells such that super-k is avoided. The occurrence of super-k is indicated directly with a particular dynamic well production measurement, called flowmeter, and is also weakly correlated with thin, high permeability facies intervals at the wells, themselves correlated to static well measurements. Our reservoir characterization is accomplished through joint conditioning to well flowmeter and static data. We achieve joint conditioning through a combined fracture / facies model, in which the static data is honored with the facies model, and the dynamic data is honored predominantly with the fracture model. The dynamic conditioning is achieved through inversion, with an optimization algorithm that iterates various fracture realizations, combined with a static facies realization, through a reservoir flow simulator, to optimize fracture parameters under the minimization of the error between observed and simulated flowmeter response. Our approach is practical in that a conventional commercial flow simulator is utilized for the forward simulation, because sources, rather than discretization, are used to model discrete fracture system flow. Furthermore, the use of sources is restricted to the forward simulation, and therefore our algorithm may be used with any discrete fracture or facies model.

2 Super-k model workflow

We hypothesize that super-k structures which are laterally extensive, for example those systems of greatest interest that enable premature production of injection water at adjacent producers over large distances, are comprised predominantly of one or more discrete fracture networks (DFN). Our reservoir model is therefore composed of a DFN model, embedded into a facies model. The facies model is a training image based model[1], composed of three facies, and is conditioned to the static well data. Stochastic realizations of the facies model may be perturbed by the probability perturbation method[2], although currently the facies model is kept predominantly frozen, due to a previous study[3] that concluded flowmeter data is significantly more sensitive to the DFN model, than to the facies model. A reservoir permeability map is derived from the facies model, with each facies assigned a constant permeability. The DFN model is object based, in which DFNs are represented by vertical 2D planes. Our optimization algorithm[4] incorporates stochastic realizations of the DFN model, in which various DFN geometry and location parameters are deformed through gradual deformation of random number sequences[5]. DFN regional density may also be varied. The DFN model is then mapped into a model composed of sources, to be later input into a commercial flow simulator, along with the facies permeability map, for the forward flow simulation. Iterations of this workflow are performed until an acceptable match of the flowmeter data is achieved, thereby defining one "history matched" model. The process is repeated to obtain several history matched models.

2.1 Super-k study area reservoir characterization data

Super-k is directly measured by means of a special production testing device, called the flowmeter, which establishes the flow contribution from individual intervals in the well. The signature of super-k is abnormally high flowmeter production rates from thin zones. The operator defines super-k as intervals in which the liquid volume flux, injection or production, exceeds 500 barrels/day/ft. The definition is general because super-k has otherwise escaped precise characterization. The 500 barrels/day/ft flux may be predicted by conventional influx performance models, but very often, this flux, and higher, cannot be predicted with such models. The operator has submitted actual flowmeter data, comprised of 18 surveys, taken irregularly over 25 years, from two water injection wells and six production wells, in the study area. Incidental to direct measurement of super-k flow, flowmeter testing has established that a sufficient, but not necessary, condition for super-k flow is the existence of thin, high permeability intervals that are capped above and below with impermeable surfaces. These permeability characteristics are correlated to well wireline log and well core measurements. Well logs from all study area wells and core data from one producing well, were also made available. Seismic data is currently not available for the study area, although recent and detailed field wide 3D seismic studies have shown the Arab-D to be extensively faulted[6], and Ghawar flow simulation studies, beginning very recently[7], have cited faults and fractures as integral Arab-D flow components. All wells in the study area are vertical, and no direct DFN measurements, such as imaging data, is available from the wells. Conceptual "data," the geologic facies model of the study area, has been provided by the operator's geologists, to be used in facies training images.

3 The use of sources to model discrete fracture flow

Conventional flow simulation of DFNs is burdened with significant obstacles. Two conventional methods are discretization of DFNs, and the dual grid model. We employ the source model. The source model is more popularly known as the conventional flow simulation well model, with sources referred to as connections.

A fine scale discretized DFN model may resolve the geometries of the networks very well, although at a severe cost of flow simulation convergence performance, simply because of the severe contrast in volumes between the fine fracture blocks, and adjacent coarse blocks. This cost currently renders discretization as generally not feasible for this use. Additionally, every new DFN realization requires a new discretization. Thus, any algorithm employing optimization methods, which must, by necessity, run numerous flow simulations, will require an automatic gridding module, currently still a research area, as part of the algorithm. The dual grid model was originally developed to simulate flow through fracture systems more appropriately defined by a continuum model, and therefore is generally not suitable for application to DFNs. A dual grid may be used to model DFNs by selectively assigning high permeabilities, in the fracture grid, to those blocks which contain fractures. However, since the dimensions of the fracture grid is identical to the matrix grid, resolution of fracture geometries is poor, or at least limited to the coarse grid resolution. The dual grid is also computationally expensive.

The conventional source model, in the context of fracture flow network modeling, is represented simply as a set, S^w, of connection transmissibilities, T_j^w, constrained to be in hydrostatic equilibrium, and constrained to a zero net source volume flux,

$$S^w = \left\{ T_1^w, T_2^w, ..., T_j^w : \rho_{fluid}; q = 0 \right\},$$

where w is a fracture name, and j is the grid block containing the connection. The condition that the connections are in hydrostatic equilibrium is represented by ρ_{fluid}, the density of the well fluid. The net source volume flux, q, is constrained to zero, as both a realistic condition, and a necessary one, since conventional well models do not allow accumulation. Another, more general, way of describing S^w is as a set of sources, with each source j having a strength represented by T_j^w, and with a unique name, w, given to indicate that members of the set are governed by the same constraints. A further constraint of the set may be added, that is, a function which maps rate and location within the well, to wellbore friction pressure gradient. This constraint comprises advanced well features in most commercial simulators. The friction gradient function is not considered in this study. However, it is recognized that fracture friction pressure gradient is not insignificant relative to matrix flow pressure gradient, in many instances. We assume it is negligible in our study area, currently, while reserving the prospect of incorporating the friction gradient function, in future work.

Advantages afforded the use of the well model are many. First, adding fractures to the flow simulation is as straightforward as adding wells. Next, no updating of the flow simulation coarse block discretization is required with a new fracture network model realization. Also, the geometry of sources is not constrained - the sources in S^w may lie along a curve, as in a production or injection well, in a plane, as in a fracture plane, in multiply oriented and connected planes, or any other geometry in 3D. Furthermore, there are common conditions in which omission of all but terminal connections does not introduce significant error in predicting the effect of flow in the fracture network on reservoir performance[4]. Next, there is no limit on the total number of sets S^ω. Also, multiple sets S^ω may occupy a single flow simulation grid block. Of course, there are computational limits on the numbers of sets S^ω. Next, the transmissibility of connections are controlled by the user. This ability is available in all commercial flow simulators and is desirable due to the variability of fracture connection transmissibility due to complex, small scale fracture geometries, as in transmissibility increases resulting from brecciation. Finally, the well model approach may be used with any discrete fracture reservoir model. The only requirement is that the discrete fracture model be enabled to be mapped onto the flow simulation grid.

3.1 Representation of the fracture flow network

A fracture flow network is the apex of an extremely complex hierarchical fracture system. This entire system, or multiple systems, may be encompassed by the areal dimensions of a typical coarse flow simulation block, as in that of this study, measuring 250m x 250m. A practical discrete fracture flow network model cannot generate individual components of the network, only the culminating DFN itself. The DFN model described in Sec. 4, generates realizations of flow networks represented as isolated or intersecting planes, which in turn are mapped as sets of connections as described in this section.

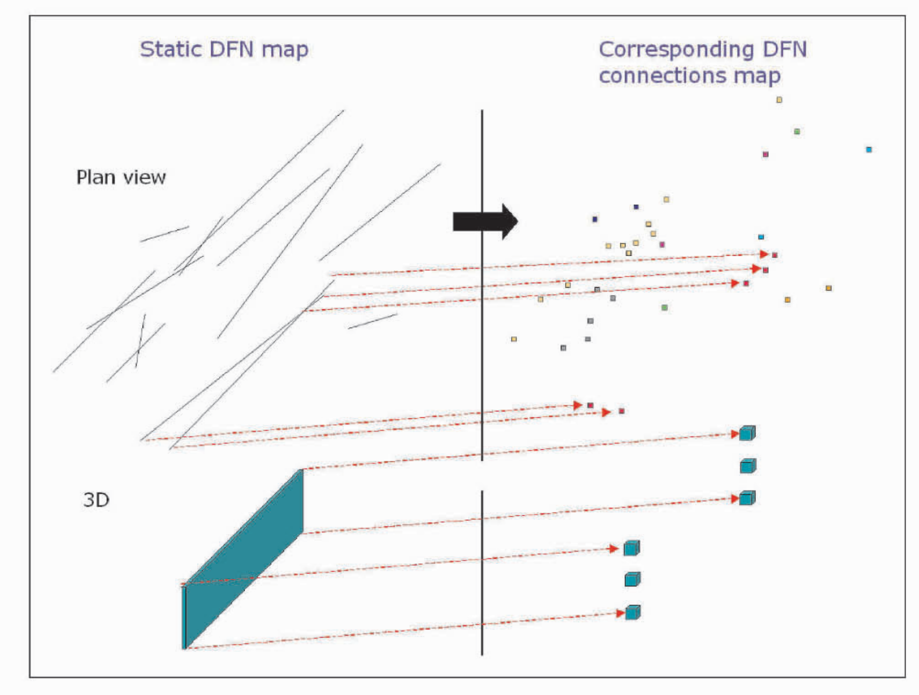


Figure 1. Mapping transmisibility connections from a DFN static model

3.1.1 *Terminal connections*

It is desirable to limit the number of well connections that are mapped to the simulation grid, as circumstances may dictate that only a few key connections are pertinent. Furthermore, a large number of low-flux connections, unnecessarily burdens computation, while not contributing significantly to the resulting effect of the fracture network flow. The flow at the terminal regions of the fracture network, for example, are often the most important to flow, and therefore most relevant to the flow simulation. Indeed, the extent of the DFN is defined by the connections of highest fracture-matrix transmissibility, separated by the greatest distances, and yet defining the basic skeleton of the network. Figure 1 presents 2D and 3D examples of connection mapping, showing the geometry of a realization of a static DFN model (Sec. 4), and its corresponding mapping of terminal connections, represented as squares or cubes. Each connection is contained in a flow simulation grid block. Figure 1 also includes fracture intersection connections. The connections are colored uniquely according to the DFN set in which they belong, that is, the flow network in which hydrostatic equilibrium is assumed. Note that in the case (bottom figure) for which a DFN possesses a vertical thickness that encompasses more than one flow simulation grid block, the entire column of grid blocks, intersected by the DFN end regions, is populated with connections.

3.1.2 *Near-well and near-neighbor connections*

Points in the network that are near well blocks may also be critical, because these are regions where fluid flux in the reservoir is large. Furthermore, the influence of fracture networks is highest when they are near wells. Also, points in the network that are near neighboring networks are important because DFNs are complex failure regions, often not confined to narrow and restricted geometries, as may be individual fractures. DFNs in close proximity may possess damage zones which overlap, thus inducing some degree of inter-DFN flow.

3.1.3 *Conditioning fracture flow networks to super-k flow intervals*

Well flowmeter data that indicates super-k, also indicates the proximity of one or more DFNs. Super-k response does not require the direct intersection of a DFN with the well in which the flowmeter data is measured[4]. The model therefore conditions a DFN to the *proximity* of the data, that is, at the well flow simulation grid block. This implies, in the case of our study for example, a proximity of 125 m.

4 An object-based DFN model with connection mapping

Figure 2 presents a rendering of an example DFN model realization. The model is constructed without a grid; the 3D grid shown is a flow simulation grid, upon which the fracture flow network model is superimposed. The grid corresponds to that used in the study: 32 x 12 x 60. The vertical exaggeration is approximately 200. The planar objects represent fracture flow networks, with varying azimuths, plunge, lengths, and vertical thicknesses. There are 28 such networks in this example. Also shown are production and injection well traces, with well connections marked with small spheres.

The cube markers in Figure 2 indicate the upper and lower extents of the vertical intervals over which flow simulation transmissibility connections are placed. For example, a pair of markers on the end of a plane, as shown in Figure 2, indicate that transmissibility connections are placed at those locations, as well as all grid blocks in a vertical line between the markers. These connection markers are placed at various points: terminal points of planes, intersection points of planes, points of intersection of planes with the grid boundary, points on the planes near to production or injection wells, and points on planes near other fracture flow network planes. These connections are shown as they are mapped onto the simulation grid, and therefore coincide with simulation grid blocks. The planes, in some instances, extend outside the boundary of the flow simulation grid. Transmissibility connections are, however, placed at points where the planes pierce the boundary, and therefore the effective terminal points of the network are at the boundary, and not beyond.

4.1 Model input

Our model stochastically distributes planes globally within the volume of the 3D flow simulation grid, according to an input total number of planes, and a regional DFN density, which may be obtained, for example, from seismic data, or from horizontal well data. The model also conditions DFNs to production or injection well intervals in which super-k flow was observed (Sec. 3.1.3). Four sets of input specify the ranges of distributions for the random drawing of length, vertical thickness, azimuth, and plunge of fracture flow network planes. The drawing currently occurs on uniform distributions within these ranges. Finally, the tolerances for which

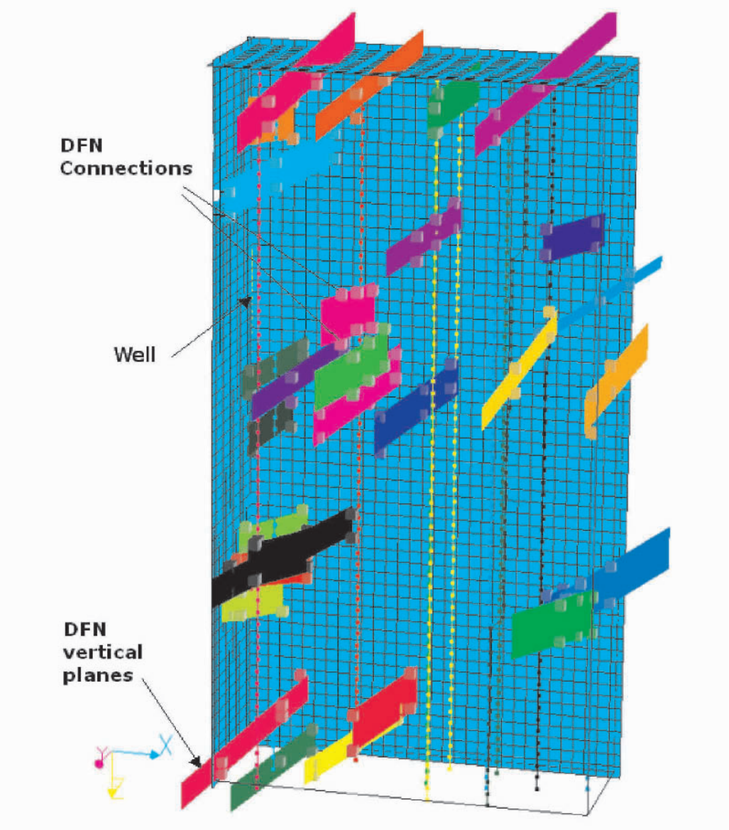


Figure 2. Elements of the DFN model

connections are placed in proximity to production and injection wells, and other fracture network planes, is input. DFN planes which approach wells or other planes within these tolerances, will have fracture connections generated in the plane, at points nearest to the object of interest.

5 Optimization study

The study area is 8 km by 3 km by 250 ft thick. There are eleven wells in the study area - nine producing wells and two water injection wells. The wells are drilled on approximately 1 km spacing, the typical spacing for the Ghawar Field. Four wells in the study area, one injector and three producers, have been identified as having exhibited super-k flow behavior. Over 25 years of study area well production performance history is available.

Our algorithm optimizes DFN parameters through minimization of simulation error relative to flowmeter data. Parameters which can be perturbed in the algorithm include: DFN density, by region, and individual DFN location, azimuth, length, vertical thickness, plunge, and connection transmissibility.

This section presents an example optimization study, to demonstrate the variation in the realizations generated, through the variation in the computed flowmeter

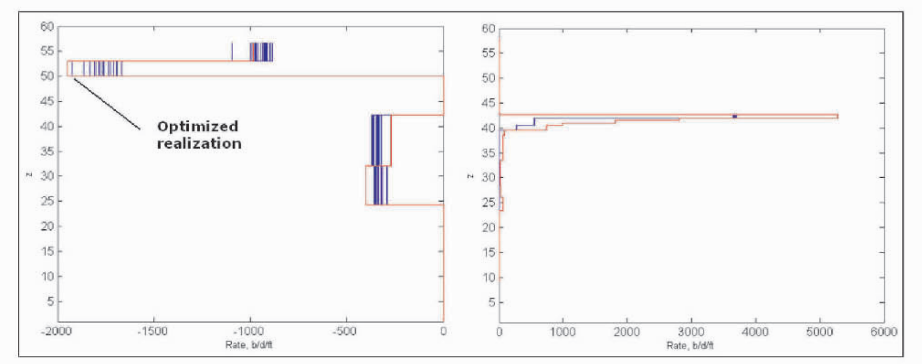


Figure 3. Optimization results for two wells. Flowmeter simulation results from 20 realizations (blue) vs. measured flowmeter data (red)

production results from flow simulations. Optimization was performed only on the locations of DFNs, in this example, although DFN azimuth, length, vertical thickness, and plunge were random, not fixed. DFN density, although varied by region, was fixed for the optimization, as was DFN connection transmissibilities. The example consists of 20 iterations. A single facies model was constructed for the 20 iterations, and subsequently, each iteration required the following key steps:

- stochastic construction of a DFN realization, under gradual deformation,
- generation of a map of connections from the DFN realization,
- imbedding of the connections map and the facies permeability map into the flow simulation grid,
- flow simulation of the combined realization,
- computation of the error of simulation vs. measured flowmeter data.

Flow simulation computes 25 years of production history from the study area, as well as flowmeter results, from eight wells. A total of 19 flowmeter surveys were conducted at various times during the producing history of the study area. Simulated flowmeter results, compared to measured data, for one survey from each of four surveyed wells will be presented here. Each flow simulation duration was 45-75 minutes. The duration of construction of each reservoir realization was a few seconds.

5.1 Optimization results

Flowmeter simulation results for two of the eight wells in which flowmeter surveys were conducted, are shown in Figure 3. These two wells possessed the most severe examples of super-k flux magnitudes. Simulation results from 20 realizations are presented for each well. The results include those from one water injection well (left), and one producing well.

The flowmeter data for each well is presented as a vertical profile of liquid flux, in barrels/day/ft. The z axis is measured in flow simulation grid cell vertical indices. The optimization error was computed in this example, for simplicity, using the flowmeter data from one well, the water injection well. Note that good simulation matches are obtained in this well, and that ultimately, an acceptable minimization of simulation vs. measured error is achieved by the optimization algorithm, as indicated in Figure 3. The producing well, displaying a super-k interval flux approaching 5500 barrels/day/ft, is not well matched, and requires further optimization. Success in optimizing a flux of this magnitude has been achieved in realizations in which a series of DFNs form effective conduits extending from the injection well to the producing well.

References

[1] Strebelle, S., Journel, A.G.: “Reservoir Modeling Using Multiple-Point Statistics,” paper SPE 71324 presented at the 2001 SPE Annual Technical Conference and Exhibition, New Orleans, Louisiana, 30 September-3 October.

[2] Caers, J.: “History matching under a training image-based geological model constraint,” SPE *Journal*, September, 2003, 498.

[3] Voelker, J.: "A Characterization of Ghawar Super-k Distribution Through Flowmeter History Updating of Training Image Based Maps," *Stanford Center for Reservoir Forecasting Report* No. 15, May, 2002.

[4] Voelker, J.: "The Use of the Conventional Well Model to Predict the Effect of Discrete Fracture Network Flow on Reservoir Flow Performance," *Stanford Center for Reservoir Forecasting Report* No. 17, May, 2004.

[5] Hu, L.Y., Blanc G., and Noetinger B.: "Gradual deformation and iterative calibration of sequential stochastic simulation," Mathematical Geology, vol. 33, no. 6, 2001

[6] Alexander, D. W.: "Impact of 3-D seismic data on reservoir characterization and development, Ghawar Field, Saudi Arabia, " *AAPG Studies in Geology No. 42*, edited by P. Weimer and T. L. Davis, eds., 1996, p. 308-317

[7] Valle, A., Faulhaber, J.J., Keith, T.H., Hsueh, P.T.: “Development of an Integrated Reservoir Characterization and Simulation Model for a Heterogeneous Carbonate Reservoir, Arab-D Reservoir, East Flank of Ghawar Field,” paper SPE 37778, presented at the 1997 SPE Middle East Oil Show, Bahrain, 15-18, March.

COMBINING METHODS FOR SUBSURFACE PREDICTION

PETTER ABRAHAMSEN
Norwegian Computing Center, PO Box 114 Blindern, NO-0314 Oslo, Norway

Abstract. The depth to subsurfaces in a multi-layer model is obtained by adding the thickness of layers. However, the choice of layering is not unique so there will often be alternative ways of obtaining the depth to a particular subsurface. Each layer thickness can be described by a stochastic model accounting for uncertainties in the thickness. Stochastic models for the depth to subsurfaces are obtained from these. Alternative layer models will give alternative stochastic models and thus alternative depth predictions for the same subsurface. Two approaches to resolve this ambiguity is proposed. The first uses an established method of unbiased linear combination of predictors. The second and new approach combines the alternative stochastic models into a single stochastic model giving a single predictor for subsurface depth. This predictor performs similarly to the approach combining several predictors while drastically reducing computational costs. The proposed method applies to layered geological structures using a combination of universal or Bayesian kriging and cokriging.

1 Introduction

Consider the problem of mapping the depth to subsurfaces separating geological layers within a petroleum reservoir. The top and base of the reservoir are often visible on seismic data so accurate depth maps are obtained from depth converted travel time maps. The internal layering will rarely exhibit reliable seismic signals, so the thickness trend of each layer is mapped using geological interpretation of bore-hole data. The total thickness of the internal reservoir layers will seldom add up to the thickness depicted from seismic data. This ambiguity must be resolved to provide consistent depth maps describing the reservoir layers.

Two approaches for resolving this ambiguity is discussed. The first approach is adapted from econometrics and forecasting (Bunn, 1989; Granger, 1989), and consists of predicting the depth to the subsurfaces by combining alternative depth predictions ‘in an optimal manner’. This approach works, but it is computationally expensive. An alternative and new approach is therefore proposed. Instead of combining the predictors, different stochastic models are combined. The result is a single stochastic model with a single associated predictor. These approaches perform very similar but computer expenses are dramatically reduced.

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 601–610.

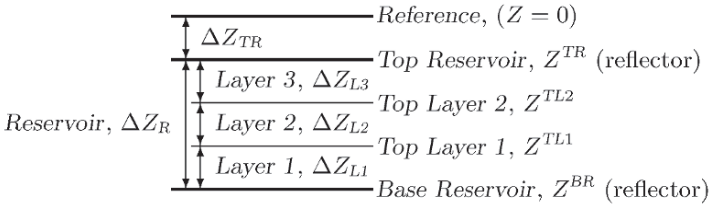


Figure 1. Schematic illustration of a reservoir formation. The double arrows indicate a stochastic model for the thickness of the corresponding layer, ΔZ_i.

2 Position of the problem

The stochastic model for the thickness of layer i may include a deterministic trend, $m_i(\mathbf{x})$, and a zero mean Gaussian random field, $\epsilon_i(\mathbf{x})$, for the residual error (Abrahamsen, 1993):

$$\Delta Z_i(\mathbf{x}) = m_i(\mathbf{x}) + \epsilon_i(\mathbf{x}); \qquad \mathbf{x} \in \mathbb{R}^2.$$

The stochastic model for the depth to subsurface l becomes $Z^l(\mathbf{x}) = \sum_{i=1}^{l} \Delta Z_i(\mathbf{x})$.

Figure 1 shows a schematic cross-section of a reservoir where subsurfaces *Top Reservoir* and *Base Reservoir* are assumed to be seismic reflectors. For a layer i bounded by two seismic reflectors, the trend is $m_i(\mathbf{x}) = v_i(\mathbf{x})\Delta t_i(\mathbf{x})$, where $v_i(\mathbf{x})$ is velocity and $\Delta t_i(\mathbf{x})$ is the seismic travel time. Models for the depth to *Top Reservoir* and *Base Reservoir* would be $Z^{TR}(\mathbf{x}) = \Delta Z_{TR}(\mathbf{x})$ and $Z^{BR}(\mathbf{x}) = \Delta Z_{TR}(\mathbf{x}) + \Delta Z_R(\mathbf{x})$ respectively (see Figure 1 for notation). Thickness trends, $m_i(\mathbf{x})$, for the internal reservoir layers are usually based on little data and many assumptions so the variance of the corresponding residual error could be large.

As an alternative method for obtaining the depth to *Base Reservoir* the thickness of all the internal layers could be added to *Top Reservoir*: $Z^{BR}(\mathbf{x}) = \Delta Z_{TR}(\mathbf{x}) + \Delta Z_{L3}(\mathbf{x}) + \Delta Z_{L2}(\mathbf{x}) + \Delta Z_{L1}(\mathbf{x})$. In practical applications the former model is preferred because seismic data are assumed more accurate than geological interpretation.

Lets look at a less obvious situation. The depth to *Top Layer 1* can be obtained by adding layer thicknesses to the depth of *Top Reservoir* or by subtracting layer thicknesses from the depth to *Base Reservoir*:

$$Z^{TL1}(\mathbf{x}) = \Delta Z_{TR}(\mathbf{x}) + \begin{cases} \Delta Z_{L2}(\mathbf{x}) + \Delta Z_{L3}(\mathbf{x}) & \text{add to } TR \\ \Delta Z_R(\mathbf{x}) - \Delta Z_{L1}(\mathbf{x}) & \text{subtract from } BR. \end{cases}$$

It is not obvious which alternative to choose. Since the seismic reflectors are assumed accurately determined, Figure 1 suggests that subtracting from *Base Reservoir* could be a better choice. Similar reasoning suggest that obtaining *Top Layer 2* by adding *Layer 3* to *Top Reservoir* is a good choice. However, these choices leaves a ‘gap’ between the two subsurfaces so the trend, $m_{L2}(\mathbf{x})$, for the thickness of *Layer 2* is never considered. Moreover, this choice has a serious implication on

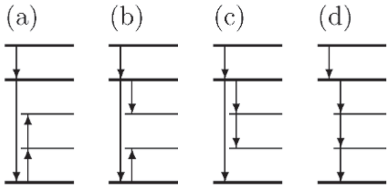


Figure 2. Four alternative methods for constructing the depth to the subsurfaces given in Figure 1.

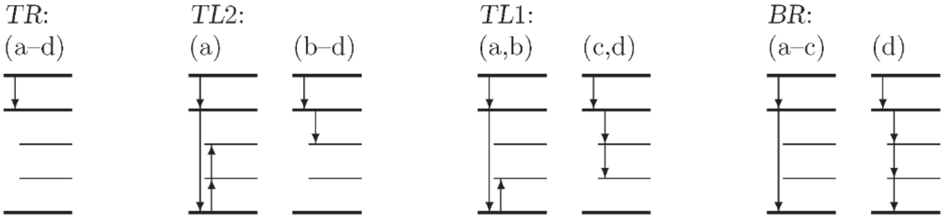


Figure 3. Alternative methods for constructing the depth to *Top Reservoir* (*TR*), *Top Layer 2* (*TL*2), *Top Layer 1* (*TL*1), and *Base Reservoir* (*BR*). Labels correspond to the graphs in Figure 2.

the uncertainty of the thickness of *Layer 2*: Assuming the residual errors, $\epsilon_i(\mathbf{x})$, to be independent implies that the variance of the thickness of *Layer 2* is

$$\mathrm{Var}\big\{Z^{TL2}(\mathbf{x}) - Z^{TL1}(\mathbf{x})\big\} = \mathrm{Var}\big\{\Delta Z_{L3}(\mathbf{x})\big\} + \mathrm{Var}\big\{\Delta Z_{R}(\mathbf{x})\big\} + \mathrm{Var}\big\{\Delta Z_{L1}(\mathbf{x})\big\}.$$

This variance is usually significantly larger than $\mathrm{Var}\big\{\Delta Z_{L2}(\mathbf{x})\big\}$ and the possible strong correlation between the depth to *Top Layer 1* and *Top Layer 2* is lost.

The discussion has motivated the need for an approach where several methods can be used simultaneously, so the unpleasant need for choosing one particular method becomes obsolete.

Figure 2 shows four graphs, each corresponding to a method for constructing *all* subsurfaces in Figure 1. The depth to a particular subsurface is found by following the arrows to the subsurface; an arrow pointing downwards means that the corresponding thickness is added whereas an arrow pointing upwards means that the corresponding thickness is subtracted. Although some graphs give the same result for a particular subsurface, the *dependencies* between the subsurfaces are different in all four graphs. This has an implication on the predictors for each subsurface (Abrahamsen, 1993). Thus, each graph in Figure 2 corresponds to a method for prediction of the set of subsurfaces so there are actually four different predictors for each of the four subsurfaces.

Figure 3 illustrates the methods for constructing subsurfaces slightly differently. Whereas, Figure 2 gives a method for *all* subsurfaces in each graph, Figure 3 shows the different methods for *each* subsurface. Each graph in Figure 3 are labelled using the labels in Figure 2. Note that although Figure 2 contains four graphs (or methods), there is only one or two possible ways of adding layers to obtain a particular subsurface.

3 Stochastic models for subsurfaces

A stochastic model with a linear trend for the thickness of layer i reads

$$\Delta Z_i(\mathbf{x}) = \underbrace{\mathbf{g}_i'(\mathbf{x})\,\boldsymbol{\beta}_i}_{m_i(\mathbf{x})} + \epsilon_i(\mathbf{x}); \qquad \mathbf{x} \in \mathbb{R}^2, \tag{1}$$

where $\mathbf{g}_i(\mathbf{x})$ is a vector of P_i known (deterministic) functions, $\boldsymbol{\beta}_i$ is a vector of P_i unknown parameters, and the residual error, $\epsilon_i(\mathbf{x})$, is a zero mean Gaussian random field specified by the standard error, $\sigma_i(\mathbf{x})$, and the correlation function $\rho_i(\mathbf{x}, \mathbf{y})$. The correlation function and standard error for the residual error can occasionally be estimated from bore-hole data. (1) is a multiple linear regression model with a correlated error term.

A typical model for the thickness of a layer i is $\Delta Z_i(\mathbf{x}) = \beta_{i1} + h(\mathbf{x})\,\beta_{i2} + \epsilon_i(\mathbf{x})$, where $h(\mathbf{x})$ is a trend supplied by geologists, so that $\mathbf{g}_i'(\mathbf{x}) = [1, h(\mathbf{x})]$. A typical travel time based model for the thickness of a layer i is $\Delta Z_i(\mathbf{x}) = [\beta_{i1} + \beta_{i2}\bar{t}_i(\mathbf{x})\,]\,\Delta t_i(\mathbf{x}) + \epsilon_i(\mathbf{x})$, where $\bar{t}_i(\mathbf{x})$ is the seismic travel time to the mid-point of layer i and $\Delta t_i(\mathbf{x})$ is the seismic travel time in layer i. So $\mathbf{g}_i'(\mathbf{x}) = \big[\Delta t_i(\mathbf{x}),\ \bar{t}_i(\mathbf{x})\Delta t_i(\mathbf{x})\big]$. A positive value for β_{i2} gives the widely encountered velocity increase at larger depthes due to compaction (Faust, 1951; Acheson, 1963). A similar velocity model was used by Hwang and McCorkindale (1994) for predicting the velocity field and by Xu, Tran, Srivastava and Journel (1992) for predicting depth. The residual error must account for the uncertainty in travel time (Walden and White, 1984; White, 1984) and the uncertainty in the interval velocity field (Al-Chalabi, 1974, 1979; Abrahamsen, 1993).

Consider now a multi-layer model including L layers and subsurfaces. The depth to the lth subsurface is

$$Z^l(\mathbf{x}) = \sum_{i=1}^{l} \Delta Z_i(\mathbf{x}) = \mathbf{f}^{l\prime}(\mathbf{x})\,\boldsymbol{\beta} + \mathcal{E}^l(\mathbf{x}),$$

where $\mathbf{f}^{l\prime}(\mathbf{x}) = \big[\mathbf{g}_1'(\mathbf{x}) \ \cdots \ \mathbf{g}_l'(\mathbf{x}) \ \mathbf{0}' \ \cdots \ \mathbf{0}'\big]$ and $\boldsymbol{\beta}' = \big[\boldsymbol{\beta}_1' \ \cdots \ \boldsymbol{\beta}_L'\big]$. Here, $\mathbf{0}$ are zero vectors replacing $\mathbf{g}_i(\mathbf{x})$ for $i = l+1, \ldots, L$, and $\mathcal{E}^l(\mathbf{x}) = \sum_{i=1}^{l} \epsilon_i(\mathbf{x})$.

4 Choice of Predictor

The best linear unbiased predictor for a random field with an unknown linear trend is the universal kriging predictor. All subsurfaces in a multilayer model are statistically dependent (covariates) since they all depend on the thickness of at least one common layer. So the kriging predictor for any subsurface should be conditioned on available depth observations from all correlated subsurfaces using universal cokriging (Abrahamsen, 1993).

The kriging predictors depend on the geometry of the observations, the choice of linear trends for the layer thicknesses, and the statistical properties of the residuals. So alternative methods, such as those illustrated in Figure 2, give different predictions for the same set of observations.

Universal kriging was used by Hwang and McCorkindale (1994) and cokriging by Jeffery, Stewart and Alexander (1996) for predicting velocity fields for depth conversion. Xu et al. (1992) finds that universal kriging and collocated cokriging give similar results for depth conversion. Using universal kriging however, gives the possibility of using non-linear relationships between depth and travel time.

5 Approaches to resolving the ambiguities

5.1 COMBINING PREDICTORS

This approach is an adaption of a method used in time series analysis and forecasting and is reviewed by Bunn (1989) and Granger (1989). The idea is to make a linear combination of alternative predictors.

For a subsurface l in Figure 1 there are four possible predictors corresponding to the four different graphs or methods in Figure 2: $Z^{*l}_{(\mathrm{a})}(\mathbf{x})$, $Z^{*l}_{(\mathrm{b})}(\mathbf{x})$, $Z^{*l}_{(\mathrm{c})}(\mathbf{x})$, and $Z^{*l}_{(\mathrm{d})}(\mathbf{x})$. A linear combination of these is a possible combined predictor:

$$\begin{aligned} Z^{*l}(\mathbf{x}) &= w^{*l}_{(\mathrm{a})}(\mathbf{x})Z^{*l}_{(\mathrm{a})}(\mathbf{x}) + w^{*l}_{(\mathrm{b})}(\mathbf{x})Z^{*l}_{(\mathrm{b})}(\mathbf{x}) + w^{*l}_{(\mathrm{c})}(\mathbf{x})Z^{*l}_{(\mathrm{c})}(\mathbf{x}) + w^{*l}_{(\mathrm{d})}(\mathbf{x})Z^{*l}_{(\mathrm{d})}(\mathbf{x}) \\ &= \boldsymbol{w}^{*l\prime}(\mathbf{x})\,\mathbf{Z}^{*l}(\mathbf{x}). \end{aligned} \tag{2}$$

Assume that each predictor is unbiased and that the covariance matrix, $\mathcal{C}^{*l}_{ab}(\mathbf{x}) = \mathrm{Cov}\big\{Z^{*l}_a(\mathbf{x}) - Z^l_a(\mathbf{x}), Z^{*l}_b(\mathbf{x}) - Z^l_b(\mathbf{x})\big\}$, of the predictors is known. Then, an unbiased predictor with the minimum prediction variance is obtained using weights

$$\boldsymbol{w}^{*l}(\mathbf{x}) = \boldsymbol{\mathcal{C}}^{*l^{-1}}(\mathbf{x})\,\mathbf{e}/\left(\mathbf{e}'\boldsymbol{\mathcal{C}}^{*l^{-1}}(\mathbf{x})\,\mathbf{e}\right), \tag{3}$$

where $\mathbf{e}$ is a vector of unit entries. This result is analogous to the weights obtained in ordinary kriging.

To form $\boldsymbol{\mathcal{C}}^{*}(\mathbf{x})$ requires the kriging prediction variances and even the prediction covariances between all predictors at any location $\mathbf{x}$. Thus, the drawback of this method is the necessity to evaluate several predictors, prediction variances, and prediction covariances for every subsurface.

5.2 COMBINING STOCHASTIC MODELS

This new approach propose that the alternative stochastic models for the depth to a particular subsurface should be combined according to the magnitude of the residual error for each model. A linear combination of the alternative stochastic models is considered.

There are two different methods and stochastic models for the depth to *Top Layer 1* in Figure 1 according to Figure 3. A linear combination reads

$$Z^{TL1}(\mathbf{x}) = w^{TL1}_{(\mathrm{a,b})}(\mathbf{x})\,Z^{TL1}_{(\mathrm{a,b})}(\mathbf{x}) + w^{TL1}_{(\mathrm{c,d})}(\mathbf{x})\,Z^{TL1}_{(\mathrm{c,d})}(\mathbf{x}). \tag{4}$$

The weights $w^{TL1}_{(\mathrm{a,b})}(\mathbf{x})$ and $w^{TL1}_{(\mathrm{c,d})}(\mathbf{x})$ are chosen to minimise the residual error variance of $Z^{TL1}(\mathbf{x})$ subject to the condition that the weights add to one:

$$\mathbf{w}^{l}(\mathbf{x}) = \mathbf{C}^{l^{-1}}(\mathbf{x})\,\mathbf{e}/\left(\mathbf{e}'\mathbf{C}^{l^{-1}}(\mathbf{x})\,\mathbf{e}\right), \tag{5}$$

where the elements of the covariance matrix, $C^l_{ab}(\mathbf{x}) = \mathrm{Cov}\big\{Z^l_a(\mathbf{x}), Z^l_b(\mathbf{x})\big\}$, are calculated using

$$\mathrm{Cov}\big\{Z^l_a(\mathbf{x}), Z^m_b(\mathbf{y})\big\} = \mathrm{Cov}\big\{\mathcal{E}^l_a(\mathbf{x}), \mathcal{E}^m_b(\mathbf{y})\big\} = \sum_{i \in \substack{\text{common}\\ \text{intervals}}} s^{ab}_i \,\mathrm{Cov}\big\{\epsilon_i(\mathbf{x}), \epsilon_i(\mathbf{y})\big\}, \qquad (6)$$

where $s^{ab}_i = -1$ when interval i is added in one model and subtracted in the other. Otherwise, $s^{ab}_i = 1$. The combined residual error variance is $\mathrm{Var}\big\{Z^l(\mathbf{x})\big\} = [\mathbf{e}'{\mathbf{C}^l}^{-1}(\mathbf{x})\,\mathbf{e}]^{-1}$ which is less than or equal to $\mathrm{Var}\big\{Z^l_a(\mathbf{x})\big\}$ for any method a.

Similar combinations must be constructed for all the subsurfaces. It is then straightforward — but requires some bookkeeping — to calculate covariances between depth observations from different subsurfaces. This leaves one stochastic model and a single associated predictor for the depth to any of the subsurfaces.

Combining predictors is based on the principle of minimising the prediction error. Combining stochastic models however, is a heuristic approach which must be justified by comparing the results to the results obtained when combining predictors. An example will illustrate that the two approaches give almost identical results. The advantage of the latter approach is speed because only one predictor for each subsurface is needed.

5.3 RELATED APPROACHES

(2) has the form of a multiple linear regression model for $Z^{*l}(\mathbf{x})$ with $Z^{*l}_a(\mathbf{x})$; a = (a), (b), (c), (d) as regressors and the weights as unknown parameters. A constant term accounting for possible bias can be added (Granger, 1989). This approach, called 'stacked regression' by Wolpert (1992) and Breiman (1992), either requires historical data or a large data set allowing cross validation. The review by Clemen (1989) compares different methods for choosing the weights in (2).

6 A Synthetic Example

6.1 STOCHASTIC MODELS

Consider the schematic cross section of a reservoir formation illustrated in Figure 1 and assume constant thickness for each reservoir zone:

$$\Delta Z_{Li}(x) = \beta_{Li} + \epsilon_{Li}(x); \qquad i = 1, 2, 3,$$

where $x \in \mathbb{R}$ since only a cross-section is considered. Moreover, $\Delta Z_{TR}(x)$ and $\Delta Z_R(x)$ are given as:

$$\Delta Z_{TR}(x) = \Big[\beta_{TR1} + \beta_{TR2}\,\big\{t_{TR}(x) - \mathrm{mean}\big(t_{TR}(x)\big)\big\}\Big]\, t_{TR}(x) + \epsilon_{TR}(x)$$
$$\Delta Z_R(x) = \left[\beta_{R1} + \beta_{R2}\frac{5-x}{10}\right] \Delta t_R(x) + \epsilon_R(x).$$

The expressions in the square brackets are the seismic velocities. A cross-section of the travel times is shown in Figure 4. A positive value for β_{TR2} gives the usual

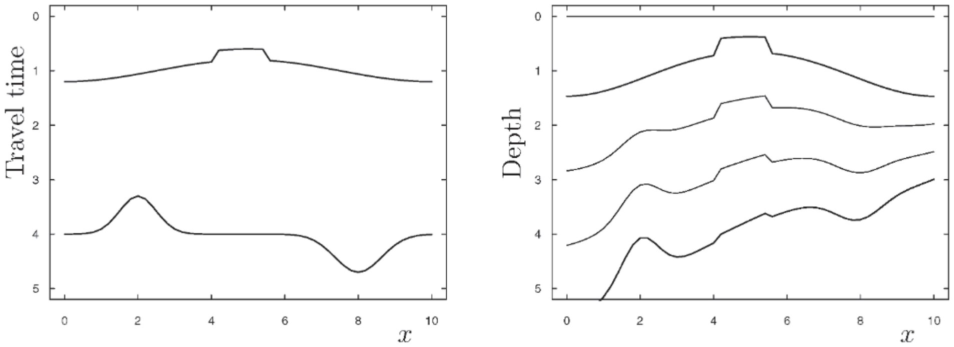


Figure 4. Travel times to *Top Reservoir* and *Base Reservoir* (left). Depth trends obtained when choosing all β's equal to one (right).

Table 1. Specified residual errors σ_a^l, calculated weights w_a^l, and combined residual errors σ^l. The methods correspond to labels in Figure 3. Note how the weights favour the assumed most accurate models.

Subsurface	Method	Res. error	Weight	Comb. res. error
l	a	$\mathrm{Var}\left\{Z_a^l\right\}^{1/2}$	w_a^l	$\mathrm{Var}\left\{Z^l\right\}^{1/2}$
Top Reservoir:	(a–d)	0.1	1	0.1
Top Layer 2:	(a)	0.316	0.31	0.194
	(b–d)	0.224	0.69	
Top Layer 1:	(a,b)	0.245	0.62	0.202
	(c,d)	0.300	0.38	
Base Reservoir:	(a–c)	0.141	0.92	0.139
	(d)	0.361	0.08	

velocity increase with increasing depth causing the subsurfaces to be more curved than the travel times. A positive value for β_{R2} leads to a reduced interval velocity for higher x values causing *Base Reservoir* to tilt upwards towards the right. The standard errors of the residual errors for the layer thicknesses are chosen as $\sigma_{TR}(x) = \sigma_R(x) = 0.1$, $\sigma_{TL1}(x) = \sigma_{TL2}(x) = \sigma_{TL3}(x) = 0.2$, and they are assumed to be independent for simplicity in the example.

When combining models, the possible methods for constructing each of the subsurfaces are illustrated in Figure 3. The covariance matrix $\mathbf{C}^l$ has dimension one for *Top Reservoir* and dimension two for the three other subsurfaces. Note that $\mathbf{C}^l$ is independent of x because the standard errors of the residual errors are assumed constant. The resulting weights from (5) and residual errors of the combined models are given in Table 6.1.

Choosing all β-parameters equal to one and combining the trends according to (4) using the weights in Table 6.1, gives the depth trends illustrated in Figure 4. This set of trends are considered the 'truth' in the following.

6.2 SIMULATION EXPERIMENTS

Universal kriging predictors split into two parts; the estimated trend and the local fitting to observations (Cressie, 1993). The estimated trend depends heavily on the trend model while the local fitting is mainly dependent on the shape of the correlation function (variogram). The estimated trend will be used instead of the full kriging predictor because using trends will exaggerate potential problems and differences between approaches. Moreover, areas away from wells are the most difficult to predict accurately and therefore the areas of the greatest concern. The conclusions reached will carry over to the less sensitive kriging predictors in areas outside well control. It is reasonable to assume that results are valid in the vicinity of well observations. Three different approaches are tested:

1. Combining predictors (estimated trends) according to (2).
2. Estimating trends using stochastic models combined similar to (4).
3. Like 2. but using a Bayesian prior on the β's.

Ten sets of depth observations have been drawn from a multinormal distribution with the expectations given by the depth trends in Figure 4, and covariances obtained from the weights and (6). The location of these observations are obtained by dividing the x-axis into three segments and "drilling" one vertical well in each segment at a random location.

Trends have been estimated for each set of observations using the three approaches. The resulting ten sets of trends are seen in Figure 5.

The first approach combines four trend estimates (see Figure 2) using weights obtained from (3). Note that these weights depend on x.

The second approach, combining models, gives some trends that are far off the 'true trends' in Figure 4. This is caused by severe collinearity making it almost impossible to estimate some of the β-parameters.

The third approach, imposing a prior distribution on the β-parameters with expectations 0.5 and $\text{Cov}\{\boldsymbol{\beta}\} = \text{diag}(2)$, dramatically improves the estimates of the β-parameters. The corresponding ten trends in Figure 5 show a behaviour very similar to the one obtained by combining predictors. The prior distribution effectively restricts the parameter space so that extreme β estimates are prohibited. Choosing a prior with large standard error (200%) and an expectation far away from the true value (0.5 compared to 1) still gives good results. So the approach is appearently robust to poor and vague prior specifications.

6.2.1 *Bias and Errors*

To investigate bias and accuracy, one hundred sets of observations have been drawn using the procedure described above. Figure 6 (left) displays the average empirical bias (difference between 'true' and estimated trend) of the resulting hundred estimated trends for subsurface *Top Layer 2*. It is seen that all three approaches have little bias (<3%). This is expected since model assumptions for the estimators agree with the model that generated the data. However the average empirical trend error, $\text{Var}\{\text{'true'} - \text{estimated trend}\}^{1/2}$, in Figure 6 (right) clearly show that the model combination approach have difficulties. The two other approaches produce acceptable empirical prediction errors.

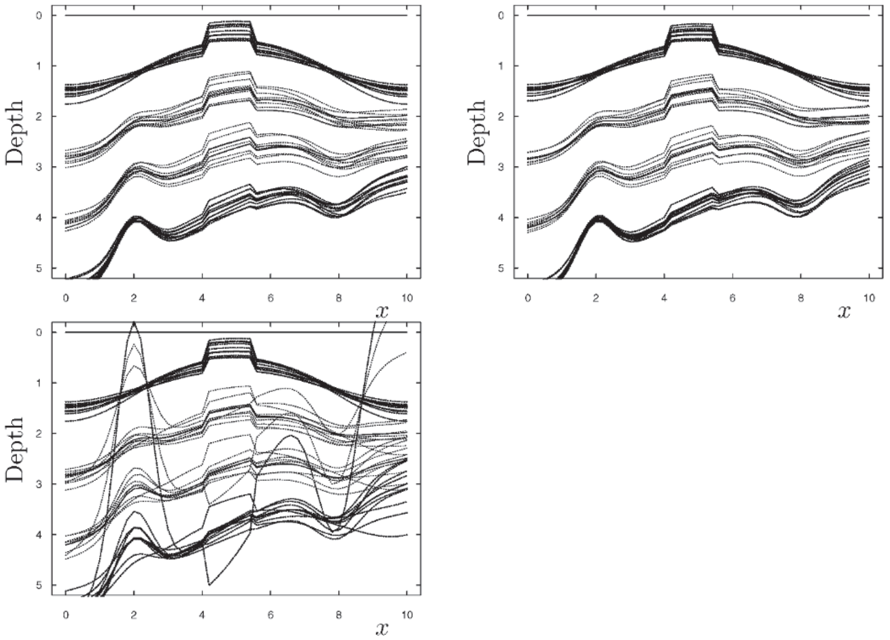


Figure 5. Ten sets of trends obtained by: 1. combining estimated trends (top left), 2. combining stochastic models (bottom left), and 3. combining stochastic models and using a Bayesian prior on the β's (top right).

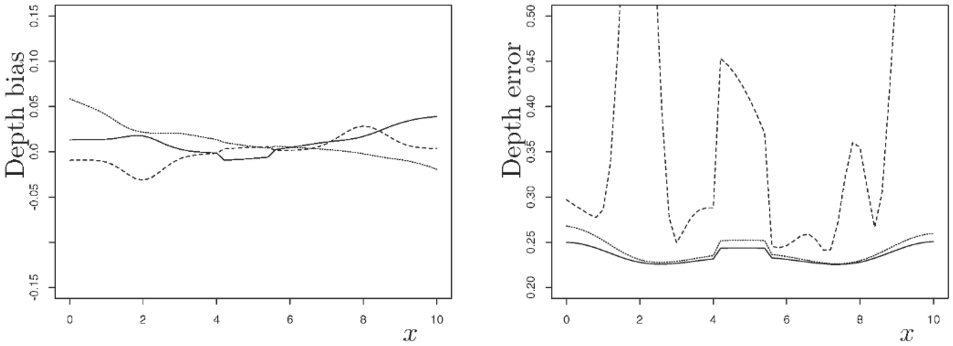


Figure 6. Average empirical bias (left) and error (right) from 100 simulations for *Top Layer 2*. ($\cdots$) combined predictors, (- - -) combined model, and (—) combined model with priors on β parameters.

7 Closing remarks

Two solutions to the problem of combining different methods for obtaining depthes to subsurfaces have been discussed. The first approach combines alternative predictors and the second approach merges alternative stochastic models. The latter approach is approximately 10 times faster but suffers from collinearities that are handled by imposing a prior distribution. The example showed that even a misspecified prior gave good results so the approach appear to be robust.

When combining predictors, a rigorous minimisation criteria for the prediction error is employed. The approach combining models however, uses a heuristic minimisation criteria for the residual variance. The usefulness of this approach is therefore justified by its performance. The two methods gave almost identical results for the synthetic example so in this situation it is possible to conclude that the model combination approach performs equally well.

The method has been implemented in commercial software and has been succesfully used in many field studies.

Acknowledgements
The work was supported by a research fellowship from The Research Council of Norway. The author is grateful to Arne Skorstad, Erik Bølviken, Magne Aldrin, and Gudmund Høst for many valuable comments.

References

Abrahamsen, P., 1993, Bayesian kriging for seismic depth conversion of a multi-layer reservoir, *in* A. Soares, ed., Geostatistics Tróia '92, proc. '4th Inter. Geostat. Congr.', Tróia Portugal, 1992, Kluwer Academic Publ., Dordrecht, p. 385–398.

Acheson, C. H., 1963, Time-depth and velocity-depth relations in western Canada, Geophysics v. 28, no. 5, p. 894–909.

Al-Chalabi, M., 1974, An analysis of stacking, RMS, average, and interval velocities over a horizontally layered ground, Geophysical Prospecting v. 22, no. 3, p. 458–475.

Al-Chalabi, M., 1979, Velocity determination from seismic reflection data, *in* A. A. Fitch, ed., Developments in Geophysical Exploration Methods—1, Applied Science Publ. Ltd., London, chapter 1, p. 1–68.

Breiman, L., 1992, Stacked regression, Technical Report 367, Dept. Statistics, Univ. California Berkeley, 15 p.

Bunn, D., 1989, Forecasting with more than one model, J. Forecasting v. 8, no. 3, p. 161–166.

Clemen, R. T., 1989, Combining forecasts: A review and annotated bibliography, Internat. J. Forecasting v. 5, p. 559–583.

Cressie, N., 1993, Statistics for Spatial Data, revised edn, John Wiley & Sons, New York, 900 p.

Faust, I. Y., 1951, Seismic velocity as a function of depth and geological time, Geophysics v. 16, no. 2, p. 192–206.

Granger, C. W. J., 1989, Invited review: Combining forecasts—twenty years later, J. Forecasting v. 8, no. 3, p. 167–173.

Hwang, L., and McCorkindale, D., 1994, Troll field depth conversion using geostatistically derived average velocities, The Leading Edge v. 13, no. 4, p. 561–569.

Jeffery, R. W.,, Stewart, I. C. F., and Alexander, D. W., 1996, Geostatistical estimation of depth conversion velocity using well control and gravity data, First Break v. 14, no. 8, p. 313–320.

Walden, A. T., and White, R. E., 1984, On errors of fit and accuracy in matching synthetic-seismograms and seismic traces, Geophysical Prospecting v. 32, no. 5, p. 871–891.

White, R. E., 1984, Signal and noise estimation from seismic reflection data using spectral coherence methods, Proc. IEEE v. 72, no. 10, p. 1340–1356.

Wolpert, D., 1992, Stacked generalization, Neural Networks v. 5, p. 241–259.

Xu, W.,, Tran, T. T.,, Srivastava, R. M., and Journel, A. G., 1992, Integrating seismic data in reservoir modeling: The collocated cokriging alternative, *in* 67th Ann. Tech. Conf. and Exhibition, Soc. of Petroleum Engineers, Washington DC, p. 833–842.

PROCESS-BASED RESERVOIR MODELLING IN THE EXAMPLE OF MEANDERING CHANNEL

ISABELLE COJAN, OLIVIER FOUCHE, SIMON LOPEZ, JACQUES RIVOIRARD

Centre de Géostatistique, Ecole des Mines de Paris, Fontainebleau - France

Abstract. The need of models for heterogeneous reservoirs has stimulated, for the last 15 years, the development of stochastic models, for instance pixel-based (indicator or truncated Gaussian simulations) or object-based (e.g. Boolean). Such models are flexible, some are easy to condition, however the geometry and arrangement of sedimentary bodies often lack realism when the geological context is better known. Multipoint statistics, for instance, are looking to improve the situation.

Yet another generation of models, both process-based and stochastic, is able to provide satisfactory modelling for heterogeneous reservoirs by reproducing the depositional processes. This is illustrated in the case of reservoirs associated to meandering fluvial systems. The model consists of: 1) a channel evolving through time either continuously (according to hydraulics equations) or discontinuously (by avulsion); 2) the consistent deposition of the different sedimentary bodies (point-bars, crevasse splays, overbank alluvium…). In order to be operational, the model depends on a limited number of parameters and is computationally quick, while being able to produce a variety of architectures. Multirealizations are available thanks to the stochastic nature of parameters. The parameters can be inferred from data (e.g. through spatial statistics such as vertical proportion curves of facies). The model can be regionally constrained (e.g. to seismic), and it allows for some conditioning to well data.

1 Introduction

The need of models for heterogeneous reservoirs has stimulated the development of stochastic models. For instance pixel-based models, such as truncated Gaussian simulations or sequential indicator simulation, are flexible and relatively easy to condition. They aim to indirectly reproduce lithofacies through their geostatistical correlations. Conversely object-based models (e.g. Boolean) directly locate geological bodies in space. However the complex geometry and arrangement of sedimentary bodies reproduced by all these models often lack realism when the geological context is better known. This explains the recent development of multipoint statistics to make use of training images, when available.

Yet another solution to produce realistic models is to combine a stochastic and a process-based approach, providing that the geological processes are known well enough. This approach has been previously investigated from a mathematical point of

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 611–619.

view in the case of sediments deposited under a water depth varying with subsidence: Jacod and Joathon (1971, 1972) have in particular made some simulations, including conditional simulations for such deposits, and Matheron (1969) has computed the stationary limit distribution of deposits in a number of typical cases. While such works have been largely ignored, the ever-increasing performance of computers now allows for generating and visualizing such process-based models relatively easily. This is what is being developed here for meandering channelized reservoirs (Lopez, 2003). A process-based stochastic approach has also been used by Hu et al. (2002) to simulate the internal geometry of deltaic sandstone bodies.

2 Description of the model

2.1 Channel evolution

The difficulty to realistically represent a meandering channel in object based models was at the origin of the present developments. However, meandering rivers have long been a subject of interest for scientists and hydraulic studies dating back to the eighties (Ikeda et al. 1981, Sun et al. 1996) have allowed the development of 2D equations that represent realistic geometries of an evolving meandering river migrating on a floodplain. More specifically these equations, which are obtained from linearization of St-Venant equations under the assumptions of constant channel width, large curvature, and steady state, describe the evolution process of the channel centreline. Starting from an initial centreline, which may be a broken line with insignificant variations, the process rapidly develops its own meandering period (Fig. 1). From time to time, the two sides of a meander connect, resulting in a cut-off and an abandoned meander. The consequent reduction of the channel length compensates for the increase of the meander lengths when the channel migrates. Because of cut-offs, the process is not reversible in time, which could have been useful for conditioning to datapoints. While the equations are deterministic, we are in the case of a pre-stochastic, or chaotic, situation. A small change in the initial state of the channel results in channels that look alike, but are located differently due to variations in the exact position of the cut-offs, just like different realizations of a stochastic process. The process has been used by Gross and Small (1998) to provide the first 3D block models. In our model the equations have been extended to 3D, to take into account the variations in local slope, notably at a cut-off location (Lopez, 2003). This process describes the continuous evolution of the channel when migrating on its floodplain, and is discretized at a time step of, say, 1 year.

However, occasionally, and preferentially where the velocity of flow in the channel is locally at a maximum, a levee breach may occur. This may then lead to a discontinuous evolution of the channel if the breach is used to change its path. This results either in a chute cut-off, when the new channel only crosscuts the outer part of a meander, or in an avulsion, when the channel takes a new path downstream. So-called “regional avulsions” take place upstream of the modeled domain.

2.2 DEPOSITS

When the channel migrates, it incises the outer side of the meanders, while depositing point bars in the inner side (Fig. 2). The succession of these sigmoid deposits form complex shapes, the connectivity of which is important, as they are usually populated with sand having good reservoir properties (Fig. 3). Where levee breaching occurs, crevasse splays are immediately deposited (Fig. 4), and possibly followed by an avulsion. From time to time, an overbank flood occurs, resulting in the deposition of sediments over the floodplain and causing the aggradation of the system (increase of its level). The granulometry and the thickness of the deposit are decreasing away from the channel (as a negative exponential, in the model). This tends to increase the difference of height between the levees (borders) of the channel and the surrounding plain, until the phenomenon is compensated by an avulsion lowering the elevation of the channel on the plain. Finally lowland deposits such as organic matter, which constitutes good geological markers, may cumulate in the lowest parts of the floodplain in the interval between two overbank floods.

2.3 EDGE EFFECTS

The model attempts to represent meandering channelized reservoirs at the scale of reservoirs, typically in a 2D rectangular area with sides of a few kilometres in length, with sloping to the East, and a single channel flowing from West to East. The output is a 3D block model, consisting of the different deposits. While the 2D area represents the domain of interest, running the processes that rule the evolution of the channel and the depositions necessitates consideration of a domain extension.

Due to migration, a channel, located initially within the domain, may for instance happen to cross the longitudinal sides of the domain, requiring a lateral extension. Similarly, a downstream extension is required to allow the channel being intersected more than once by the downstream limit of the domain. The same problem also applies for the upstream limit, but being complicated by the necessary movement permitted at the upstream extremity of the channel.

Avulsions also require an extension, for the new channel may exceed the limits while still ruling the deposition over the domain. In the case of a regional avulsion, the new channel enters the domain preferentially where elevation is minimal. This would preferentially be located at the upstream corners of the domain if there is no lateral extension. On the contrary, a large lateral extension allows the possibility of having at a time a new channel entirely outside the domain, and the whole domain entirely covered by overbank sediment.

2.4 STOCHASTIC ASPECTS

The model makes use of different sources of scientific knowledge including: physical processes, sedimentological processes, as well as a number of results and observations reported in the literature that are desirable to provide a realistic model. Practical consideration of the occurrence of levee breaches, and the shape and dimensions of crevasse splays, gives some insight into the number of parameters, whose values can be chosen to be constant or variable. One very convenient way to introduce variability in the model is by randomizing parameters. For instance, the intensity of an overbank flood (i.e. the aggradation at levees) can be taken as fixed or be randomized with a given mean. Randomness is especially helpful to generate events whose occurrence is not exactly predictable: for example random selection of the location of a levee breach among a population of channel points with local maximum velocity, or the random generation of overbank floods with a given frequency. We have previously seen that in the deterministic process for meandering channel, randomness was at hand through very small variations in the intial state. In addition, explicit randomization of parameters allows multirealizations of the model.

3 Control of the model

3.1 Parameters

While randomizing the occurrence of overbank deposits, for instance, allows for different simulations, the key parameter which characterizes the occurence of overbank floods is frequency. The number of such key parameters, which rule the essential aspects of the model, can fortunately be limited, which is a necessity for the model to be operational. They include for instance, the width and depth of the channel, the slope of the floodplain, the erodibility coefficient controling the velocity of migration, and the frequency and intensity of other elements such as overbank floods, avulsions, etc. Despite the limited number of key parameters, the model is rich enough to produce very different simulations, in terms of the amount or connectivity of sand for example, by changing one or a few parameters (Fig. 5).

3.2 Statistics

In addition to visualization, statistics such as the mean proportions of facies or vertical proportion curves, can help the practitioner to choose or to modify parameters from data.

3.3 Regional conditioning

Migration is proportional to the erodibility coefficient. This can be taken as constant over the whole domain, or made to vary, either as a deterministic function or as a regionalized variable. Given such an erodibility map, the channel will confine its behaviour to areas that are more or less erodable (Fig. 6). This is thought to be very useful in order to take into account the information provided, for instance, by a seismic time slice. Moreover, if the map changes in a continuous manner, like different times of a seismic block, the channel will adapt.

3.4 CONDITIONING AT WELL DATAPOINTS

Conditioning at well data points is a very difficult and challenging issue. Oliver (2002) for instance, proposes to move and distort a non conditional channel to make it go through data points. Due to the number of elements of our model and to their evolution in time, we prefer conditioning from inside the model, acting on the evolution processes of the channel itself, namely migration and avulsion. Regional conditioning through an erodibility map was a first illustration. To favour the migration of the channel towards a datapoint where point bar must be deposited, we locally increase the erodibility map using a geostatistically simulated correction (Fig. 7). However if the distance is too large, we will first use an avulsion to approach the channel. In our model, the different types of sediments are deposited one upon the each other, with the exception of sand or mudplug that are deposited where the channel has previously eroded, and which are the only replacement facies available. It follows that the channel must keep away from data points where overbank shales, for instance, must be deposited before and after this deposition. Erodibility is then used to prevent the channel from going through such a point.

This engineering approach to conditioning does not yield to the theoretical conditional model. Such a theoretical model is not available and moreover, if it were, it would implicitely give a confidence to the assumed unconditional model itself that is not guaranteed in practice. The described conditioning through the evolution processes and the erodibility is still under devlopment, but seems to be a flexible approach and is able to honour several wells together. A 3D data management allows for the selection of active data to be used at a given time for conditioning.

4 Conclusion

When processes are known, process-based stochastic models allow for the representation of realistic geometries and arrangements of different geological sets, as illustrated in the case of meandering channelized reservoirs. The limited number of key parameters to be chosen or inferred, as well as the fast computation, allow the model to be operational. Yet, it can produce a variety of architectures by varying these parameters. Due to the stochastic aspect, multiple realizations can be provided for a given specified model.

Soft regional conditioning with seismic, or hard conditioning at well data points is a chalenging problem. However practical solutions are being designed by controlling the process itself, and that may give acceptable approximations for conditional simulations.

Non conditional simulations can otherwise be used as training images (or blocks) for geologists or engineers. This is all the more interesting in that pictures of actual systems or outcrops may provide a biased view of what is to be obtained after further erosion and sedimentation.

Such a new generation of process-based stochastic models can be developed in other systems, e.g. fluvial (multi-channel, braided or anastomosing rivers), deep sea (turbidites, Das (2002)), or carbonates.

Acknowledgments

GDF (Gaz de France), SHELL, and IFP (French Institute for Petroleum) are greatly acknowledged for their support through the Meandering Channelized Reservoirs Consortium.

References

Das, H. S., 2002, Numerical modeling of submarine channel morphology, Ph. D. Thesis, University of South Carolina.

Gross, L.J. and Small, M.J., 1998, River and floodplain process simulation for subsurface characterization, *Water resource research*, vol 34(9), p. 2365-2376.

Hu, L.Y., Joseph, Ph. And Dubrule, O., 2002, Random genetic simulation of the internal geometry of deltaic sandstone bodies. SPE 24714. In Proc. 1992 SPE annual technical conference and exhibition, p. 535-544.

Ikeda, S., Parker, G. and Sawai, K., 1981, Bend theory of river meanders. Part 1. Linear development, *Journal of Fluid Mechanics*, vol. 112, p. 363-377.

Jacod, J. and Joathon, P., 1971, Use of random-genetic models in the study of sedimentary processes. Math. Geol. 3, no. 3, p. 219-233.

Jacod, J. and Joathon, P., 1972, Conditional simulation of sedimentary cycles in three dimensions. In: Merriam D.F. (ed) Proceedings of the International Sedimentary Congress, Plenum Press.

Lopez, S., 2003, Modélisation de reservoirs chenalisés méandriformes, approche génétique et stochastique, PhD, Ecole des Mines de Paris. www.cg.ensmp.fr/Chenaux

Matheron, G., 1969, Les processus d'Ambarzoumian et leur application en géologie. Technical report N-131, Centre de Morphologie Mathématique, Ecole des Mines de Paris.

Oliver, D. S, 2002, Conditioning channel meanders to well observations, Math Geol, vol. 34, no. 2, p. 185-201.

Sun, T., Meakin, P., Jossang, T. and Schwartz, K., 1996, A simulation model for meandering rivers, *Water resources research*, vol. 32(9), p. 2937-2954.

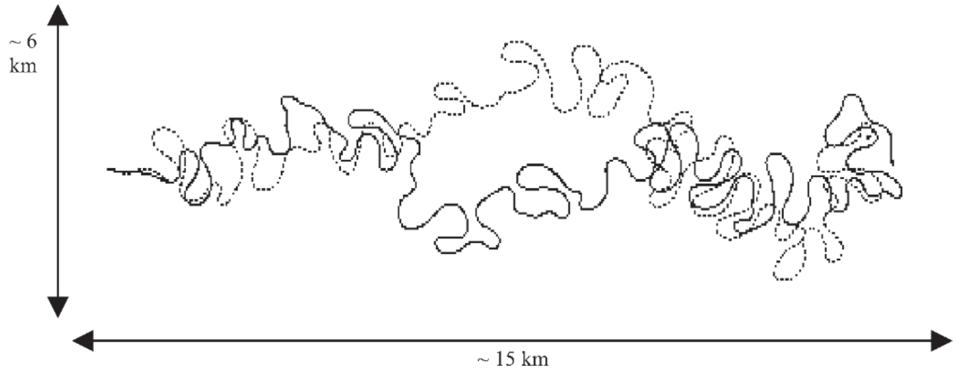


Fig. 1: Evolution of two channel centerlines after several thousands iterations, starting from quasi-straight lines from left to right which only differ by micro-perturbations.

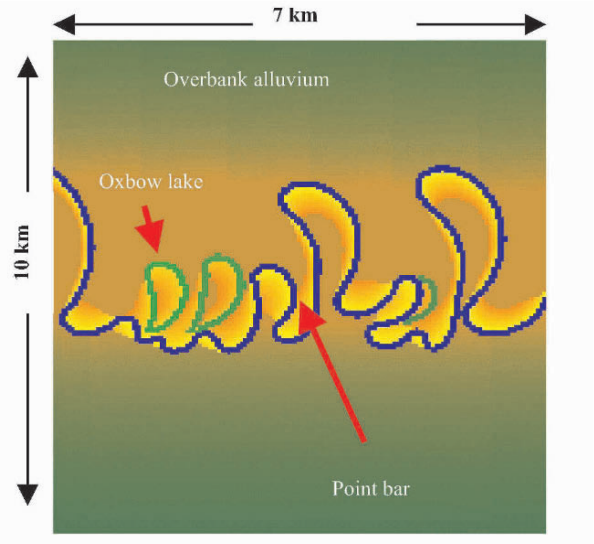


Fig. 2: Channel migration and overbank deposition (red to yellow: sands from older to more recent; dark green to light green: shales from older to more recent). Flow from left to right. From Lopez (2003).

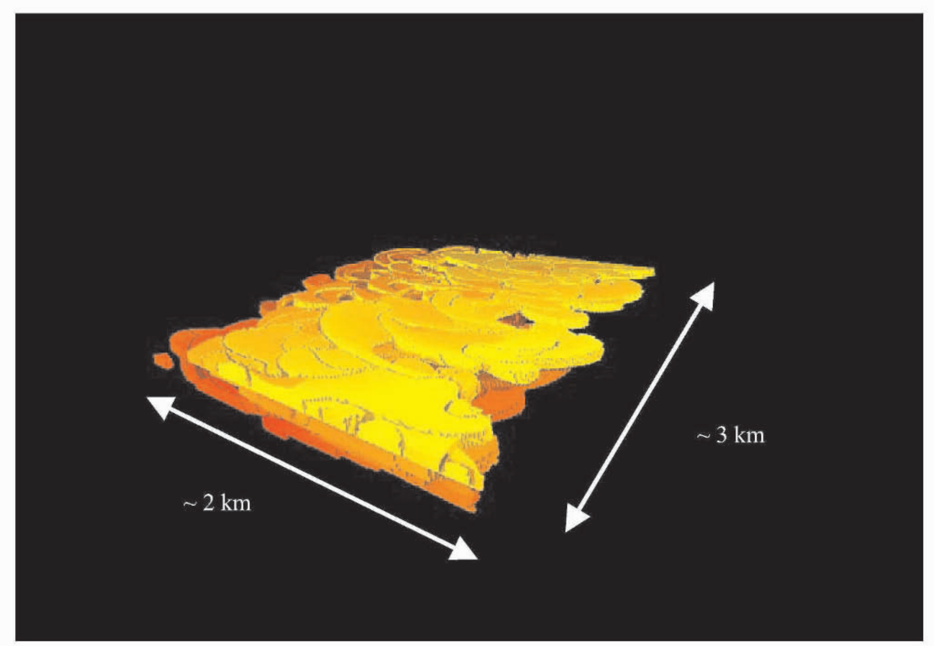


Fig. 3: Scorched view of point-bars over 20 ka (see Fig. 2 for colors). From Lopez (2003).

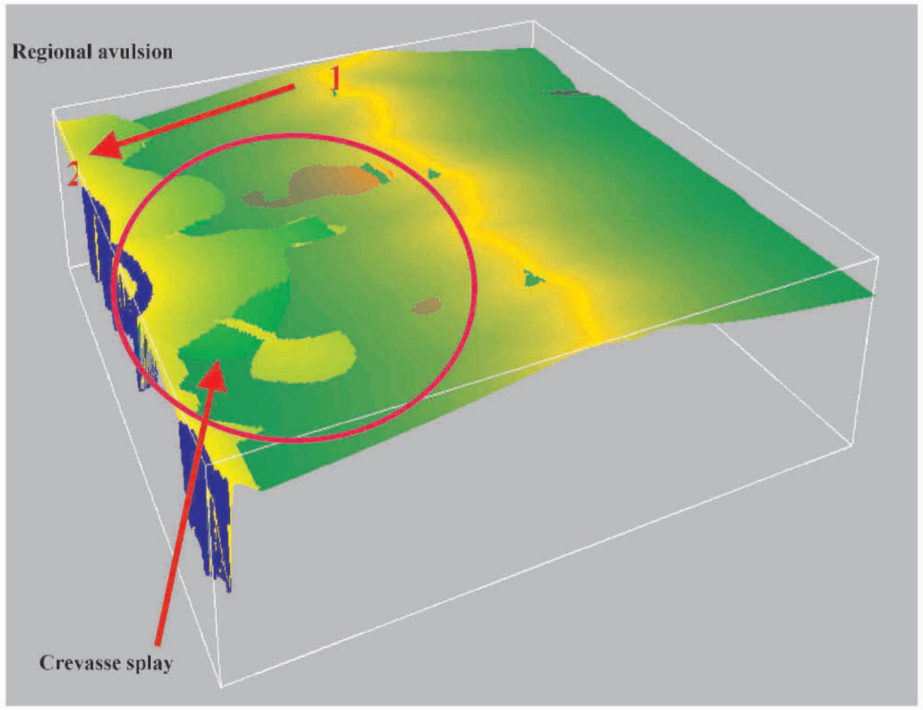


Fig. 4: Aerial view of the floodplain, including a regional avulsion followed by a levee breach with deposition of a crevasse splay (colors as in Fig. 2). From Lopez (2003).

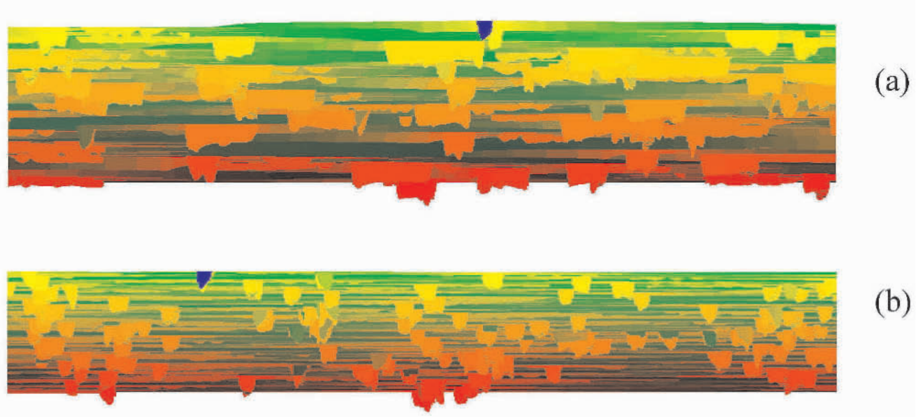


Fig. 5: Cross-sections showing different architectures produced by varying the frequency of avulsions. (see Fig. 2 for colors): (a) rare avulsions; (b) frequent avulsions. From Lopez (2003).

This page intentionally blank

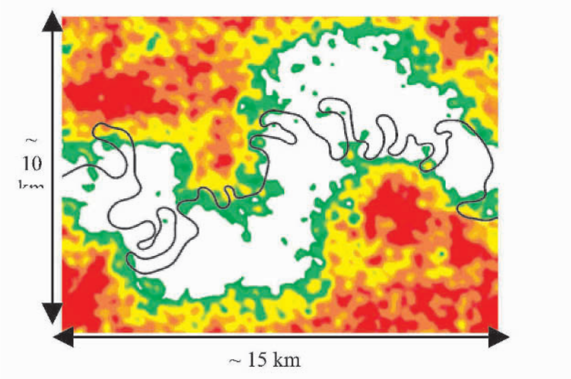


Fig. 6: A channel, initially a quasi-straight line from left to right, after 10000 migrating iterations on an erodibility map (in white: high erobility)

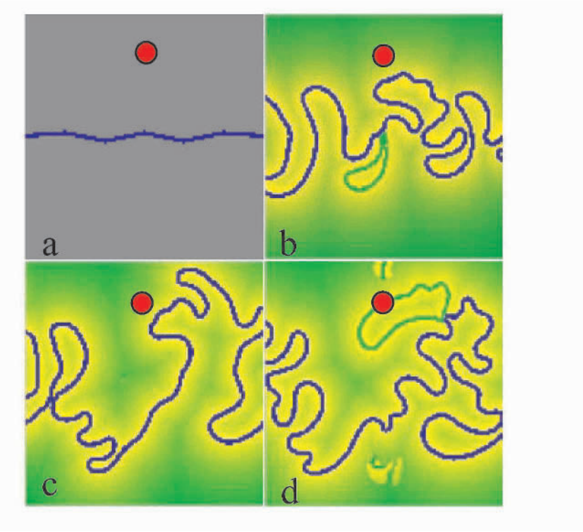


Fig. 7: Conditioning at a well datapoint: a) initial situation; the datapoint consists of point bar or mudplug; b) erodibility is used to attract the channel to the well ; c) when the channel reaches the well, deposition of point-bars is possible; d) however fixing the channel at the well location results in cut-off and mud-plug deposition, if desired.

MULTIPLE POINT GEOSTATISTICS : OPTIMAL TEMPLATE SELECTION AND IMPLEMENTATION IN MULTI-THREADED COMPUTATIONAL ENVIRONMENTS

ALVARO E. BARRERA, JOSEPHUS NI and SANJAY SRINIVASAN
Department of Petroleum & Geosystems Engineering,
University of Texas at Austin, USA TX 78712-0228

Abstract.

Sequential Indictor Simulation is a classic stochastic simulation approach that is pixel-based and utilizes kriging/cokriging to obtain estimates of the necessary conditional distributions and generate various analogous reservoir realizations of the target reservoir. However, reproduction of complex 3D patterns is not possible using traditional two-point statistics (variogram) based approaches. Therefore, new stochastic simulation approaches based on multiple point statistics have surfaced in the literature.
Since computational efficiency is a significant consideration in implementation of multiple point statistics based modeling approaches, parallel computational techniques and multithreading have to be implemented in order to render the process efficient for field scale reservoir characterization. The implementation of such a multi-threaded, multiple point simulation algorithm is discussed. The selection of an optimal spatial template is critical for capturing and reproducing complex spatial patterns observed in analogous systems. An efficient template optimization algorithm is also presented.

1. Introduction

Geostatistics has become a widespread tool for reservoir modeling and uncertainty assessment. Multiple equiprobable reservoir models constrained to data of different types and volume supports (geologic, geophysical and production data) can be generated while preserving the apparent geological structure. Traditional geostatistics is anchored on two basic concepts: the variogram model as representation of the spatial heterogeneity or continuity (more statistical than geologic), and kriging for spatial interpolation. Variogram-based geostatistics is mathematically consistent and convenient (mathematical representation of physical concepts), and its application is appropriate considering the lack of conditioning information in typical reservoir modeling scenarios and given the computational limitations facing the earlier generation of earth modelers.

Practitioners and particularly geologists have found the variogram suitable to describe geological heterogeneity within a single facies, but too limiting in describing and reproducing more organized and sharp geological features such as faults, fractures, and

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 621–630.

facies distributions among others. These geological features usually have the largest impact on the flow response. The traditional variogram-based geostatistics fail to appropriately retain the required conditional information to reproduce these features; generating amorphous realizations that exhibit maximum entropy instead of systematically organized structures and patterns as expected from prior geological knowledge. These geological features call for a different approach utilizing multiple point statistics instead of variogram models (two point statistics) to capture the required conditional information to generate fields that exhibit lower entropy, more structural organization and preserve reservoir heterogeneity.

This approach requires a training image, a numerical representation of the spatial law and an explicit non-conditional conceptual description of the geological structures and patterns to be reproduced in the field. Training images can be obtained from outcrops and photographs. Although it seems easier to define a variogram model than develop training model depicting the critical reservoir heterogeneity, however, it is important to realize that variogram models are not any less subjective or constraining, and moreover, they are more limiting and less intuitive in describing geological heterogeneity. Patterns to be reproduced can be explicitly depicted in the training model as compared to the hidden higher-order statistics implicit within traditional variogram-based geostatistical models.

The general objective of this paper is to implement a new simulation approach that is based on geological feature identification and reproduction using a unique growth-based simulation algorithm. Features are grown starting from conditional data locations based on multiple point statistics inferred using optimized spatial templates. The simulation is implemented on a multi-threaded computational environment and consequently can be used to efficiently simulate 3D models comprising of over 10 million cells. The robustness of the simulation methodology and the reproduction of realistic geological features depend on the optimality of the selected spatial template used to capture the pattern characteristics. A unique approach for optimization of the spatial template is also presented.

2. Literature Survey

Stochastic simulation was introduced by Matheron (1973) and Journel (1974) to correct for the smoothing effects and other artifacts of kriging, allowing the reproduction of the spatial variance predicted by the variogram model. Different algorithms were developed including sequential simulation (Journel, 1983, Isaaks, 1990; Srivastava, 1992; Goovaerts, 1997; Chiles and Delfiner, 1999), which has become the workhorse for many current geostatistical applications. Stochastic simulation provides the capability to generate multiple equiprobable realizations, giving birth to the idea of assessing spatial uncertainty (Journel and Huijbregts, 1978).

Stochastic simulation stepped away from the variogram and kriging based core for the first time with the Boolean object-based algorithms, introduced by Stoyan, Kendal and Mecke (1987), and Haldorsen and Damsleth (1990), in an attempt to reproduce geological features, like channels and fractures, by fitting parametric shapes. Initially

Srivastava (1992), and later Caers (1998) and Strebelle (2000), proposed the idea of borrowing conditional probabilities directly from a training image, allowing the use of higher order or multiple point statistics to reproduce geological structures and patterns. Advantages of this approach over the Boolean object-based method include the pixel-based non-iterative process and the ease of data integration of different types.

Training images and multiple point statistics methods remained largely untried until developments in computational capabilities and multiple point scanning techniques took place. The search tree, a training image scanning technique proposed by Strebelle (2000), has allowed for wider application of the multiple point statistics approach. In the search tree approach, the training proportions are recorded in a dynamic data structure that renders it convenient to store nested information such as number of outcomes of data events (joint probabilities). However, only "exact" data events directly found in the training image can be stored and later retrieved as the conditioning information in the estimation. A second approach allows the reproduction of "similar" data events by modeling and interpolating between found data events. This second approach was supported by the classification methods proposed by Arpat (2003) and Breiman et al. (1984); and neural net based models.

Currently, the basic components that characterize the application of multiple point geostatistics remain the major focus of continuous research and development efforts, with the objective of defining a more consolidated and practical methodology. These components include, among others, the generation or acquisition of numerical spatial representations to be used as training images; the optimal definition of scanning templates to capture the proper conditional information from the training image; and development of computational schemes that render the application more practical and convenient. The two last components are addressed in this paper.

3. Proposed Methodology

3.1 FEATURE IDENTIFICATION AND SIMULATION

The purpose of the multiple point geostatistics approach is borrowing conditional proportions (joint probabilities) from a numerical representation of the spatial law (training image) describing the random function. Training images require some prior information. When this prior information is uncertain, alternative training images can be used to narrow the range of uncertainty and a Bayesian approach can be adopted to incorporate that uncertainty in resultant geological models (Liu et al., 2004). The training image requires the effort of geologists to generate one or more numerical representations of spatial heterogeneity, reflecting prior information about complex shapes and geometrical patterns.

In order to capture the relevant details of reservoir heterogeneity, the size of the training data as well as the size and complexity of the spatial template can be large. Consequently the process of scanning and storing multiple point conditional probabilities can have a rather high computational requirement. Taking advantage of recent advances in computational technology using multiple *cpu* processors, the task of scanning training images is performed using multiple processing threads. The training image is segmented into multiple domains and the task of scanning each sub-domain using the specified spatial template(s) is assigned to a dedicated *cpu* thread. At the end of the scanning process, the statistics computed by each thread is compiled into a single statistical summary for the entire training image.

The objective of the scanning process is to obtain the joint probabilities:

$$\text{Prob}\{Z(\mathbf{u}_i) = z_k \forall i = 1,..,N; k = 1,...,K\} \qquad (1)$$

N is the number of nodes in the template, $\mathbf{u}_i$ is the position of the i$^{\text{th}}$ node, z_k is the k$^{\text{th}}$ threshold of the random variable $Z(\mathbf{u})$. During the simulation phase, the spatial template is placed at a location $\mathbf{u}_o$. Some nodes $\mathbf{u}_j, j = 1,..,N'$ where $N' \subset N$ is the subset of template nodes, already have simulated values. The requirement is of the probability of a simulation event $A(\mathbf{u}_o)$ conditional to the pattern $A(\mathbf{u}_j, j \in N')$ in the surrounding nodes of the template. The joint probability expression in Expression (1) can be used to calculate the requisite conditional probabilities:

$$\text{Prob}\{A(\mathbf{u}_o) \big| A\big(\mathbf{u}_j\big), j \in \big(N'\big) \subset \big(N\big)\} = \frac{\text{Prob}\{A(\mathbf{u}_o) \cup A(\mathbf{u}_j)\}}{\text{Prob}\{A(\mathbf{u}_j)\}} \qquad (2)$$

In terms of the notation in Expression (1), the numerator in Expression (2) is simply the joint probability: $\text{Prob}\{Z(\mathbf{u}_i) = z_k \forall i \notin N'; Z(\mathbf{u}_j) = z_{k'} \forall j \in N'; k, k' = 1,...,K\}$. The denominator is the joint probability in the numerator summed over all occurrences of patterns $A(\mathbf{u}_o)$:

$$\text{Prob}\{A(\mathbf{u}_j)\} = \sum_{A(\mathbf{u}_o)} \text{Prob}\{A(\mathbf{u}_o) \cup A(\mathbf{u}_j)\} \qquad (3)$$

Knowing the conditional probability given by expression (2), a simulation pattern can be obtained by drawing from the conditional probability distribution.

<u>*Remark*</u>: The simulation event $A(\mathbf{u}_o)$ in traditional multiple point statistics implementations consists of a single point event i.e. the outcome at the central node of the template given the multiple point event in the surrounding nodes. In the implementation presented here, the simulation event is allowed to be a multiple point event. Thus both the conditioning and simulation events are multiple point events. This renders the simulation process to be fast and more important, the simulated patterns exhibit better continuity.

Following scanning, the process of stochastic simulation is commenced where the inferred multiple point histogram is used in conjunction with reservoir specific data distributions to obtain realizations with realistic spatial heterogeneity. In the traditional implementations of multiple point geostatistics, the simulation nodes are visited along a random path and the probability of the central node given the configuration of pattern in the surrounding nodes is obtained from the search tree that contains a catalog of the scanned patterns. At the beginning of the simulation when only a few nodes have assigned values, considerable *cpu* time may be spent searching for unusual or infrequent patterns. In order to alleviate this problem, patterns or structures are grown from data locations in the method implemented in this study. This approach improves the continuity of the simulated patterns while reducing the artifacts in the simulated image.

Dividing the simulation domain into regions based on the density of conditional data, the conditioning data locations are visited along a random path. A node surrounding the data location is marked as "simulatable" based on whether the spatial template centered at that location contains at least a single conditional data. Corresponding to a randomly picked conditioning data location, a "simulatable" node in the vicinity of that data location is also randomly picked. The multiple point simulation event $A(\mathbf{u}_o)$ is picked from the conditional probability distribution (Expression (2)) at that location. After all the conditioning data locations are visited, the list of "simulatable" nodes is updated and the next node is selected from this updated list. The simulation is continued until all the simulation nodes have been assigned a value.

3.2 PROGRAM IMPLEMENTATION

The program is implemented in Java to facilitate cross platform compatibility. There are several inherent performance hindrances with the Java platform, however. The biggest hindrance to achieve equal performance to C++ is that of memory allocation and garbage collection. De-allocation of memory in Java is done automatically by an automated "garbage collector." Because almost everything in Java is an object, the creation of objects can be expensive, because creating an object requires calls not only to its own constructor, but also its parent class's constructor. For these two reasons, much of the reusable data types in this implementation are first created by Object Oriented design and then optimized using basic data types. Much attention was paid to choosing thread-safe data types. Other methods of optimization such as inline methods and reducing in-loop instructions and instantiations were used in conjunction with previously mentioned optimizations. Currently, the second most costly operation is an object creation step that is looped through nearly every single simulated node. This object creation overhead can be reduced by introducing a library system for objects with check-in, checkout feature. The most costly operation is repeating a search operation on the Vector data type. This problem can be alleviated by reducing the usage of such operations or by rewriting the library manually. The more detailed implementation algorithm is described below.

For the scanning process, the template generates a marginal offset at the edges of the training image in order to operate within the bounds. This marginal offset is defined by the template configuration and has an important impact in the required size of the training image when large-scale templates are used. The scanning process is described by the following steps:

1. Select a cell location in the training image and superimpose the central node of the template at that location.
2. Capture the cell values corresponding to the template nodes
3. Identify the pattern and store it. The pattern is stored in a long integer data type, I, by converting the N individual template node values, v_i, to integers in the range [1-9]

and looping over all the template nodes with the formula: $I_i = I_{i-1} + v_i * 10^{(i-1)}$; for i = 1,2,…,N; and $I_0 = 0$.

4. Repeat steps 1 through 3 until the center of the template has been located in all training image cells excluding the marginal offset.

As the training image is being scanned, each obtained pattern integer representing a data event is put into an array that is large enough to hold all possible occurrences for that training image. The array is sorted and scanned for number of occurrences of each data event. The occurrences are stored in an alternate array at the same index as the first occurrence of the data event. Both arrays are then compressed by eliminating the repetitions and zeros.

In order to optimize the performance and utilization of computational resources, the scanning process is parallelized. The first degree of parallelization is multithreading. Multithreading is much more preferable to forking a process due to much lower computing overhead in thread allocation. By applying multithreading, the scanning process time was reduced significantly – to a few seconds, by only two threads. This means that further degrees of parallelization are not required for this process including automated thread control to spawn more threads according to number of physical processors present.

The conventional stochastic simulation algorithms can be rather slow and cumbersome since they only simulate one point at a time. The scanning process defined above allows the simulation algorithm employed here to simultaneously simulate all cells under the template. All simulation nodes except the conditioning data locations are assigned a negative one (-1) value initially. Next, during the process of simulation:

1. Select a cell as center cell of the spatial template
2. Detect the pattern *t(n)* on the nodes of the spatial template by matching only non-negative values against the ones previously recorded during the scanning phase. Only when all these positive values and their positions are completely matched, then the record is chosen. Retrieve the subset of scanned patterns that exhibit the pattern t(n).
3. Take the probability of each pattern in the subset and construct a step CDF with probability on y-axis and pattern index number on x-axis.
4. Randomly sample from that CDF.
5. Get the corresponding pattern by its index and fills the cells under the template that are marked as empty.

In order to speed up the simulation process, the simulation domain is divided into a number of regions, each corresponds to a thread task. The conditioning data points are also divided into the simulation sub-domains, with each set containing a subset of simulated cells and a subset of candidate ("simulatable") cells, which are defined as uninformed cells located near previously informed cells. If any of the cells in a template during simulation is within the border of the image but beyond the quadrant border defined by the thread, the value is put down into that cell. The procedure is safe and will not conflict with another thread's execution since this portion of the code is synchronized between the threads.

3.3. RESULTS

The following are slice one through four of the simulation image, which is of size 100x130x10. The training image consists of a 3-D volume made of fluvial channels trending in the North-South direction. Conditioning data along vertical wells are assumed.

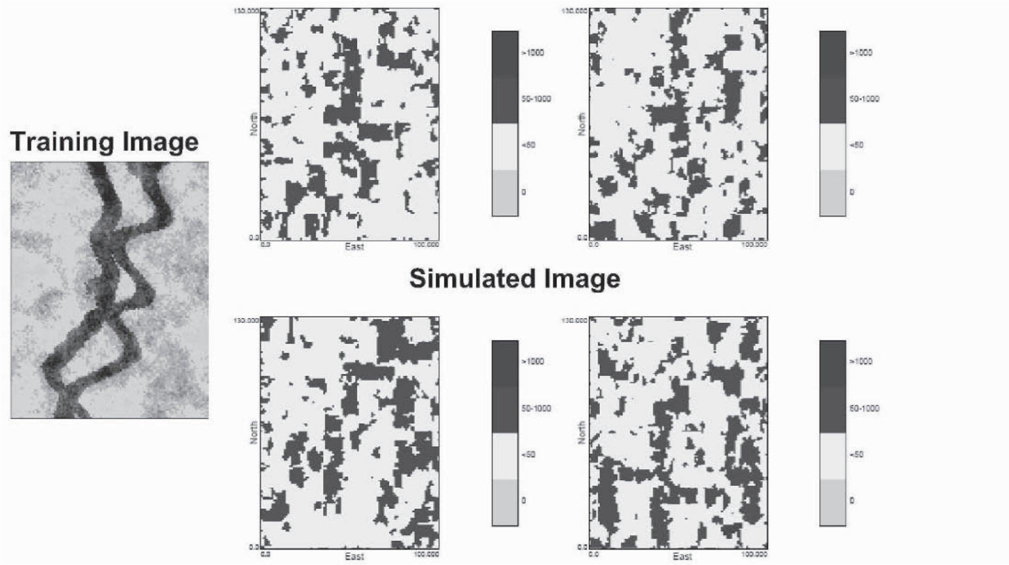


Figure 1. Slices of a 3D simulation image obtained by application of the mp statistics based algorithm. The training image (Left) consists of channels trending in the North-South direction.

The single thread operation took 2566881 milliseconds, where the multithreaded operation with four concurrent threads took on average 688000 milliseconds, which is nearly four times faster. This means that our application scaled very well with the increased number of processors present. It should be noted however, that the operating system in use is Redhat Linux 7.3 with kernel version 2.4.20 and that this kernel does not have a hyperthreading optimized scheduler, thus, the performance could be further enhanced by upgrading to a more modern kernel version.

It is also to be noted that for this particular case, the best simulation images resulted with 2-D templates. This is because the channels change in orientation from one layer of the training image to the next. This causes the 3-D templates to introduce much noise into the simulation. Despite the 2-D templates used, the slices exhibit correlation from one slice to the next due to the profusion of conditioning data in the vertical direction (due to the presence of vertical wells). Further optimization of 3-D templates to account for vertical variations will be attempted in the future.

3.4. TEMPLATE SELECTION

A suitable scanning template is essential whose size and geometry define the search neighborhood and the detection of multiple point data events describing the spatial law. Consider a grid $\mathbf{u} \in (\Theta \subset \Omega)$ where Θ is the size of the grid and Ω is the size of the simulation domain. A stationary attribute $Z(\mathbf{u})$ is assumed over the grid. Suppose a spatial template of size N is desired. The objective is to identify a template $t(\mathbf{u}_i), i = 1,..,N$ within the grid Θ such that the selected template optimally represents the dominant pattern of reservoir heterogeneity. The size and the scale of the template N are user-specified and are dependent on the available *cpu*, complexity of the geological image etc. Designating the central node in the grid Θ as $\mathbf{u}_o$, the covariance $C(\mathbf{u}_j, \mathbf{u}_o)$ between pairs of nodes $\mathbf{u}_j \in \Theta : \mathbf{u}_j \neq \mathbf{u}_o$ and $\mathbf{u}_o$ is calculated on the basis of the particular training image. Under stationarity, the required covariance is calculated by translating a two-point template $\mathbf{h}_{jo} = |\mathbf{u}_j - \mathbf{u}_o|$ over the training image. The $\Theta - 1$ covariance values $C(\mathbf{u}_j, \mathbf{u}_o), j = 1,..,\Theta - 1$ are ranked and the top N values and the corresponding locations $\mathbf{u}_i, i = 1,..,N$ define the optimal spatial template $t(\mathbf{u})$. Spatial templates at multiple scales can be obtained by choosing the grid Θ of different resolution (multiple grids), accounting for geological patterns at different scales.

In order to test the efficacy of the template optimization algorithm, multiple training images were generated exhibiting similar features but with variations in spatial orientation and anisotropy. Some examples of optimal templates obtained for these training images comprised of modified geological feature (lens) rotated at different angles are shown in Figure 2.

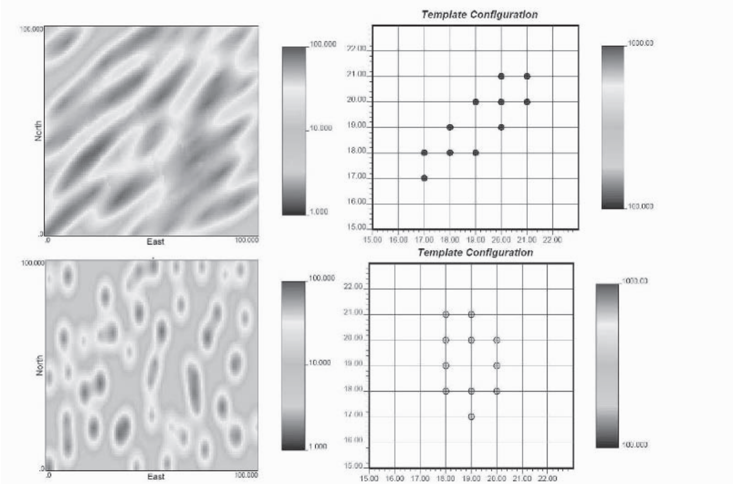


Figure 2. Template configurations obtained with the Template Selection Algorithm for similar geological features (lens) with different orientations and anisotropy.

3.4.1 *Impact of template geometry on mp statistics captured by the scanning process.*

In order to evaluate the results of the template selection algorithm and the importance of the template selection, a sensitivity study was performed. In this study, a training image exhibiting ellipsoidal lens with North-South orientation (Figure 3) was generated and scanned with six templates of the same size but different geometry. All the templates were produced by the template selection algorithm, considering different numerical representations of lens with the same size but different orientations.

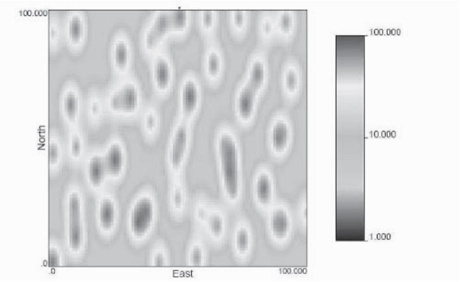


Figure 3. Training image exhibiting North-South ellipsoidal lens.

The study objective is to demonstrate the importance of an appropriate template for capturing more information about the multiple point statistics describing a particular geological feature. Hence, the multiple point statistics inference process was performed using a variety of spatial templates. It can be noticed in Figure 4, enhanced performance is obtained corresponding to Template 1, which has a North-South orientation. This template captured more relevant data events (with frequency higher than 10) from the training image (approximately 11% more than the second best Template) and consequently, the total number of occurrences retained as multiple point conditional information is increased. The difference in configuration between Templates 2 and 3 is the location of a single node; however this single node causes the number of relevant data events to drop in approximately 11%.

4 Conclusions

A unique stochastic simulation approach based on growth of objects within the simulation domain is presented. The object growth is controlled by the multiple point statistics inferred on training images. In order to render the simulation computationally efficient, the process is implemented on a multi-threaded environment. The scanning as well as the simulation processes both take advantage of multiple cpus. The computational process is demonstrated to scale well when going from 1 to 4 processors.

The selection of an optimal spatial template for retrieving the multiple point statistics is an important aspect of the proposed simulation algorithm. A fast and robust approach to derive optimal spatial templates is presented. The results show that the number of template nodes and their geometry influence the robustness of the retrieved statistics.

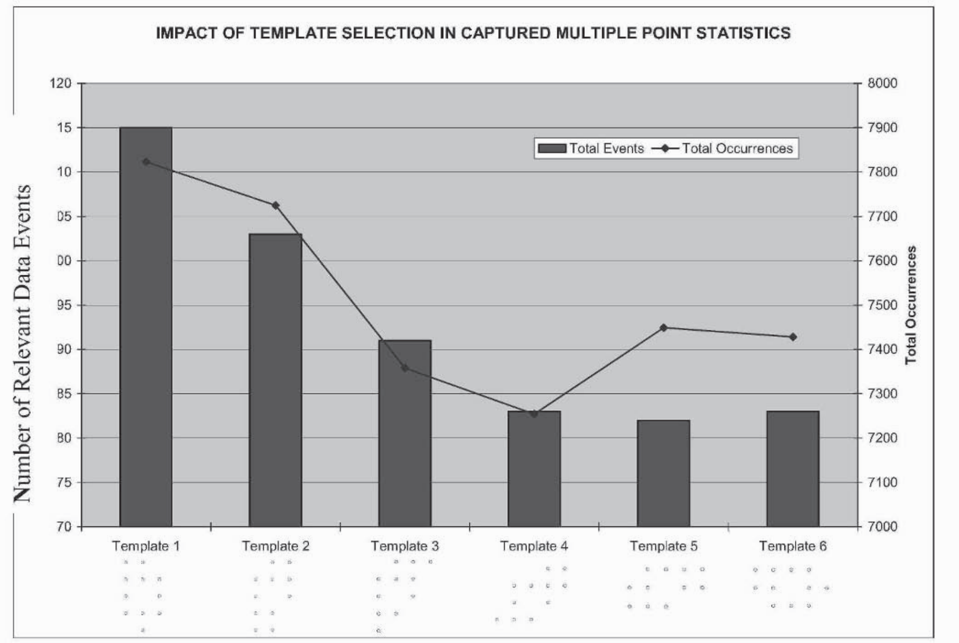


Figure 4. Number of relevant data events (frequency greater than 10 in training image) plotted against the template configuration.

5 References

Arpat, B., A pattern recognition approach to multiple-point simulation, Report no.16, Stanford Center for Reservoir Forecasting, Stanford University, 2003.

Breiman, L., Friedman, J., Olshen, R. and Stone, C., *Classification and regression trees*, publ. Wadworth, Monterrey, 1984.

Caers, J., Stochastic simulation using neural networks, Report no. 11, Stanford Center for Reservoir Forecasting, Stanford University, 1998.

Caers, J. and Journel, A.G., Stochastic reservoir simulation using neural networks trained on outcrop data, SPE paper no. 49026, 1998.

Chiles, J.P. and Delfiner, P., *Geostatistics: Modeling spatial uncertainty*, publ. Wiley, N.Y, 1999.

Goovaerts, P., *Geostatistics for Natural Resources Evaluation*, publ. Oxford Press, N.Y., 1997.

Haldorsen, H. and Damsleth, E., Stochastic modelling, *J. of Pet. Technology*, April 1990, pp.404-412, 1990.

Isaaks, E., *The application of Monte Carlo methods to the analysis of spatially correlated data*, Ph.D. thesis, Stanford University, 1990.

Journel, A.G., Geostatistics for conditional simulation of orebodies, in *Economic Geology*, vol. 69, pp. 673-687, 1974.

Journel, A.G., Non-parametric estimation of spatial distributions, *Math. Geology*, vol. 15, no. 3, pp. 445-468, 1983.

Liu, X. and Srinivasan, S., Field Scale Stochastic Modeling of Fracture Networks – Combining pattern statistics with geomechanical criteria for fracture growth, In this volume.

Matheron, G., The intrinsic random functions and their application, *Advances in Applied Probability*, vol. 5, pp. 439-468, 1973.

Srivastava, R.M., Reservoir characterization with probability field simulation, SPE paper no. 24753, 1992.

Stoyan, D., Kendall, W. and Mecke, J., *Stochastic geometry and its applications*, publ. Wiley, N.Y., 1987.

Strebelle, S., Conditional simulation of complex geological structures using multiple-point statistics, *Math. Geology*, vol. 34, no. 1, pp. 1-21, 2002.

DIRECT ASSESSMENT OF UNCERTAINTY USING STOCHASTIC FLOW SIMULATION

JULIANA Y. LEUNG and SANJAY SRINIVASAN
Department of Petroleum & Geosystems Engineering, University of Texas at Austin, U.S.A., 78712

Abstract. The main objective of this paper is to develop a technique for direct assessment of reservoir flow performance uncertainties. This is achieved via a single stochastic flow simulation that combines a model for local geologic uncertainty with a simple analog for the flow transfer function to estimate the joint probability distribution that characterizes the global uncertainty in flow performance. Our method provides the framework to go directly from local uncertainty, derived using simple spatial interpolation techniques, to flow uncertainty, skipping the intermediate steps of stochastic sequential simulation and not requiring any fine-scale flow simulations.

1 Introduction

Uncertainties in prediction of reservoir flow performance result from insufficient information available to model the reservoir and incomplete understanding of the flow processes taking place in the reservoir. Geostatistics provides a framework for incorporating data from diverse sources into the reservoir-modelling process, while realistically representing the uncertainty stemming from incomplete information.

The reservoir property at each location is modelled as a random variable (RV); the probability distribution function characterizing this RV represents the uncertainty of the attribute value at that particular location. Furthermore, the spatial distribution of the property value at all locations within the reservoir is modelled as a spatial random function (RF) that is characterized by a multivariate, joint probability distribution (Bu and Damsleth, 1996; Lia et al., 1997). The spatial distribution of reservoir attributes depicted by the numerical, geological model is generally obtained conditioned to the available information and after implementing a suitable technique for constructing and sampling from the multivariate joint distribution. This reservoir model is subsequently subjected to a flow transfer function to evaluate the flow response and production potential. Therefore, the uncertainty in the spatial distribution of reservoir attributes propagates to uncertainty in estimates of flow response. As we treat some of the reservoir properties as RVs, the governing equations for fluid flow become stochastic partial differential equations (SDEs). There are generally two basic approaches to solve these SDEs: the statistical moment equation approach (Sabelfeld and Kolyukhin, 2003) and the Monte Carlo simulation (MCS) approach. The MCS methodology is most-widely adopted for assessing uncertainties in reservoir flow performance. It involves

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 631–641.

generating multiple fine-scale depictions of the reservoir heterogeneity, each honouring the data available from diverse sources, and these images are subjected to a typical flow transfer function (a reservoir flow simulator) in order to characterize the flow performance uncertainty (Journel, 1994).

As opposed to the MCS approach, our objective in this paper is to combine the tasks of geostatistical simulation and flow modelling into one single step. Recognizing that uncertainty in flow performance comprises essentially of geologic uncertainty that is subsequently updated to global uncertainty due to flow, this paper explores avenues to directly merge the local uncertainty distribution derived using a spatial interpolation method such as kriging with a simplified model for modelling the reservoir flow performance. First, we propose a new analog technique for numerical flow simulation. Then we develop a technique to directly assess uncertainty by utilizing a single stochastic flow simulation that integrates local geologic uncertainty with our simple flow transfer model to estimate a joint uncertainty distribution that characterizes flow uncertainty. The resultant algorithm is computationally inexpensive and easy to implement. Moreover, it permits analysis of important decisions such as optimal well placement and reservoir management strategy, without having to resort to tedious stochastic simulation and expensive flow modelling. Despite its simplicity, our algorithm provides important information such as water cut and breakthrough uncertainties that are crucial for reservoir management applications. Tasks associated with uncertainty assessment, such as (1) assessment of the worth of information, (2) delineation of reservoir zones for location of future wells, and (3) ranking of reservoir models, can be reliably accomplished using the proposed method.

2 A New Analog for Numerical Flow Simulations

2.1 Background

The complete description of mass transfer in porous media requires evaluating the relative contributions of diffusion and convection. The usual way of assessing these contributions is to assume that these two effects can be decoupled and yet additive. Define the total mass flux $\mathbf{n}_i$ as the mass of species i transported per area per unit time relative to some fixed coordinates, $\mathbf{v}_i$ as the average velocity of species i, and c_i is the local concentration of species i:

$$\mathbf{n}_i = c_i(\mathbf{v}_i - \mathbf{v}_o) + c_i \mathbf{v}_o = \mathbf{j}_i + c_i \mathbf{v}_o \qquad (1)$$

where $\mathbf{v}_o$ is the convective reference velocity, $\mathbf{j}_i$ represents the diffusive flux, whereas $c_i\mathbf{v}_o$ represents the convective flux. Combining the above equation with mass balance and Fick's law, we obtain the following equation, where D is the diffusivity:

$$\frac{\partial c_i}{\partial t} = D\Delta c_i - \nabla c_i \mathbf{v}_o \qquad (2)$$

Again, the two terms on the right-hand side refer to diffusion and convection, respectively. Solutions are usually in two standard forms: (1) comprised of a series of error functions or related integrals, or (2) in the form of trigonometric series. These two types of solutions can be obtained by the method of reflection and superposition (Crank 1975) or alternatively, using the traditional method of separation of variables. The important characteristics of the solution methods are: (1) the solution corresponding to any complex boundary conditions can be constructed via the principle of superposition or summations of other elementary solutions; (2) the form of solution is generally in terms of exponential functions; diffusion influence declines exponentially in both the spatial and temporal domain. The equation governing the pressure diffusion in reservoir is similar to the mass diffusion equation and is characterized by the same exponential decay characteristics. Based on these observations, we postulate that movement of particles can be modelled as summation of dipole influences between pairs of nodes within the domain. Sudaryanto and Yortsos (2000, 2001) have successfully implemented a similar idea to optimize fluid displacements in porous media. In their approach, fluid displacement is expressed as a superposition of the response of individual wells. They have shown that the algorithm works well for both homogeneous and heterogeneous media.

Our fast analog technique utilizes particle counts as a surrogate measure of flow performance. There are two basic underlying concepts: (1) Particle movement can be decoupled into a convective term, influenced strongly by the heterogeneity of the permeability field, and a diffusive term that is governed by the gradient of concentration or pressure. (2) The particle count at a location can be obtained as the superposition of the influence exerted by all locations in the vicinity of that location. As seen in Eq. 2, mass transfer depends on parameters such as diffusion coefficient, the physical distance between two locations, and concentration or pressure difference. The dipole interaction between pairs of locations is postulated as a function of the permeabilities, distance between the pair, and difference in particle counts at the two locations. Our formulation will also include a parameter representing the normalized covariance between the two nodes to incorporate both convective and diffusive influences in one simple model.

Heterogeneity or geologic structure is described using the concept of geo-bodies in our approach. A geo-body is a group of connected blocks sharing a specified reservoir property or characteristics. Within a geo-body defined using a high permeability threshold, convection tends to dominate due to the similarity in underlying heterogeneity structure. In contrast, convection across geo-bodies is minimal due to discontinuity in the heterogeneity structure; nonetheless, fluid transport across such heterogeneities may occur due to pressure diffusion, resulting in movement of fluid particles outside and between geo-bodies. Since Fickian processes exhibit an exponential decay in space and time, our formulation will utilize an exponential covariance structure for modelling the diffusive component of fluid transport.

The notion of sources and sinks is extended in the following manner. At the initial time step, the injection wells act as sole sources and the production wells as sinks. In the injection case, the well source exerts a dipole influence on the neighbouring nodes, and the range of the influence is governed by the covariance structure. At subsequent steps, all locations that have registered an increase in particle count act as secondary sources that begin to exert influence on all locations in their neighbourhood. In the case of depletion to a producing well, all locations that have registered a decrease in particle count act as secondary sinks. The flow of fluids in the reservoir as a function of time is modelled by sequential updating of the particle count map, taking into consideration the influence of all new secondary sources or sinks.

2.2 Model Framework

Based on the preceding discussion, the following form of the dipole influence is postulated in our model: Let $w_{ij} = \overline{k_{ij}} / r_{ij}$, where $\overline{k_{ij}}$ is geometric average of permeability at nodes *i* and *j*, r_{ij} is the distance between nodes *i* and *j*.

For injection influences:

$$(I_{ij}^{n+1})_{injection} = \left(w_{ij} \Big/ \sum_{\substack{j=1 \\ j \neq i}}^{N} w_{ij} \right) \left[C(\mathbf{h}_{ij}) \times \left(P_j^n - P_i^n \right) \right], \text{ for } \left(P_j^n - P_i^n \right) \geq 0 \qquad (3)$$

$$= 0 \text{ , for } \left(P_j^n - P_i^n \right) < 0$$

For withdrawal influences:

$$(I_{ij}^{n+1})_{withdrawal} = \left(w_{ij} \Big/ \sum_{\substack{j=1 \\ j \neq i}}^{N} w_{ij} \right) \left[C(\mathbf{h}_{ij}) \times \left(P_j^n - P_i^n \right) \right], \text{ for } \left(P_j^n - P_i^n \right) \leq 0 \qquad (4)$$

$$= 0 \text{ , for } \left(P_j^n - P_i^n \right) > 0$$

where I_{ij}^{n+1} is the influence function between nodes *i* and *j* at time step *n*+1, $C(\boldsymbol{h}_{ij})$ is the covariance weighting function, P_i^n and P_j^n are particle counts at node *i* and *j* at time step *n*, respectively, and *N* is total number of grid blocks. Since $\overline{k_{ij}}$ and r_{ij} are in different units, we need to normalize the w_{ij} by its sum in order to ensure dimensional consistency and conservation of particle count in the system. The expressions (3) and (4) can be observed to be analogous to the steady state solutions for fluid flow in porous media (Darcy's Law). The particle count at node *i* at time step *n*+1 is computed as:

$$P_i^{n+1} = P_i^n + \left[\sum_{j=1}^{N} I_{ij}^{n+1} \Big/ \sum_{i=1}^{N} \sum_{j=1}^{N} I_{ij}^{n+1} \right]_{inj.} \times \Delta P_{inj} - \left[\sum_{j=1}^{N} I_{ij}^{n+1} \Big/ \sum_{i=1}^{N} \sum_{j=1}^{N} I_{ij}^{n+1} \right]_{with.} \times \Delta P_{prod} \qquad (5)$$

where P_i^{n+1} is particle count at node *i* at time step *n*+1, ΔP_{inj} is the number of particles injected during Δt, ΔP_{prod} represents particles produced during Δt. The subscripts *inj* and *with* refer to the injection and withdrawal influences, respectively. The standardization term in the denominator results in particle conservation.

If i and j belong to different geo-bodies, the covariance weighting function $C(\boldsymbol{h}_{ij})$ would follow an isotropic, exponential covariance structure; this represents the diffusive component. If i and j belong to the same geo-body, the covariance weighting function would take the larger of either an exponential covariance value (representing diffusion) or a value based on the structure of the permeability field (representing convection).

2.3 Case Study

Consider a 2-D 50x50 reservoir ($\Delta x = \Delta y = 100$m, and $\Delta z = 0.3$m) with a permeability field (in mD) shown in Figure 1a. Variogram for the geologic model is as follows: spherical structure, azimuth angle = 45^o, maximum and minimum ranges are 1800m and 600m, respectively. A range of 115m is used for the diffusive exponential structure. The results at the end of 400 days obtained from our fast analog algorithm and those obtained from flow simulator (ECLIPSE) for a single producer or injector in the middle of the reservoir, with a flow rate of 250 rm^3/d, are shown below.

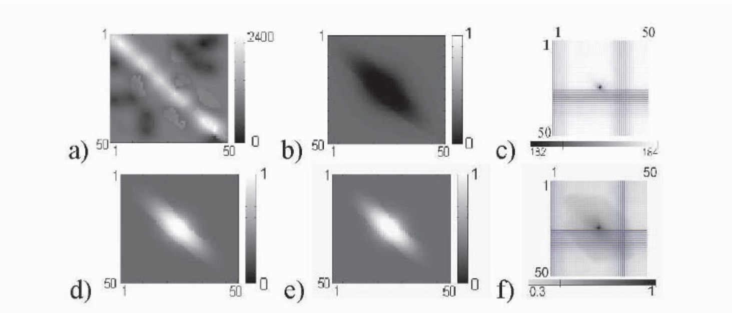


Figure 1. (a) Permeability field (in md) for the example case. (b) Particle saturations for a single producer obtained using our analog model with $\Delta t = 20$ days. (c) Water-in-place map for a single producer obtained using ECLIPSE with $\Delta t = 5$ days; the scale goes from 182.12m^3 to 184.04m^3. (d-e) Particle saturations for a single injector obtained using our analog model (d: $\Delta t = 20$ days, e: $\Delta t = 50$ days). (f) Water saturation map for a single injector obtained using ECLIPSE with $\Delta t = 5$ days.

In our methodology, the source strength is directly proportional to the total volume of injection that in turn is influenced by the time step size Δt. The larger the time step size, the stronger the source strength, and consequently the particle counts over a larger region of the reservoir get updated. Evidently, the smaller the step size, the more accurate the results are. Nonetheless, it appears that even a step size of 20 days or 50 days give reasonably good results. Moreover, it can be noticed that the results predicted by our model resembles those of a piston-type displacement, i.e. corresponding to favourable mobility ratios. A value of 115m was chosen as the isotropic diffusive range as it allows convective and diffusive influences to be manifested adequately. Heterogeneity influences increase in relative magnitude when the diffusive range is too small, or they become overly suppressed when the diffusive range is too large.

3 Application to Direct Uncertainty Assessment

3.1 Model Framework

We have presented in the previous section a fast analog technique that decouples the convective and diffusive components of the mass transfer process. The particle count at individual locations is obtained by superimposing the influence of neighbouring locations. The superposition process re-introduces the dependence between the two modes of mass transfer. In this section, the probabilistic analogy to the decomposition and superposition processes is discussed. In the probabilistic framework, decoupling amounts to independence and implies that the posterior distribution $P(A|B,C)$ is proportional to the product of the elemental probabilities $P(A|B)$ and $P(A|C)$:

$$P(A \mid B,C) \propto P(A \mid B) * P(A \mid C) \tag{6}$$

where A is flow response, such as particle counts, B is convective influence, and C is diffusive influence. The above formulation suggests that the probability of obtaining a certain flow response A given the convective and diffusive influences is a product of the two conditional probabilities. Convection is driven by heterogeneity; therefore, $P(A|B)$ can be derived using a spatial interpolation procedure such as indicator kriging that utilizes the covariance model describing the permeability heterogeneity, whereas $P(A|C)$ can be derived assuming an isotropic exponential covariance structure described in the previous section. The superposition step is an integral aspect of our fast transfer function formulation. Superposition in a probabilistic sense implies that the flow (particle count) uncertainty at each time step at a particular location is updated based on the uncertainty at surrounding nodes. The traditional, intermediate step of sequential simulation where the local uncertainty distribution derived by kriging is updated to a joint uncertainty is thus skipped. Therefore, if we consider $P(A|B)$ to be local uncertainties and $P(A|B,C)$ to be flow uncertainties, this technique provides the mechanism for updating local uncertainties into global flow-based uncertainties directly.

Prior to describing the model framework, Figure 4 (left) shows the entropy profile at a few well locations as a function of time obtained by performing a traditional Monte Carlo analysis. Entropy, a general uncertainty-measure on random variables introduced by Shannon (1948), is an excellent normalized measure of the spread of any given probability distribution p. It is defined as

$$entropy = \sum_{i=1}^{n_t} - p(z_i) * \log[p(z_i)] \tag{7}$$

where $i = n_t$ (total number of thresholds), z_i = attribute value at threshold i, $p(z_i)$ = the corresponding probability density value at threshold $z_{i.}$. A high entropy value indicates a wide distribution and larger uncertainty, while low entropy value indicates a narrow distribution and lower uncertainty. Multiple realizations of the permeability field were generated and subsequently processed through a flow simulator. It is observed from flow simulation results that entropy remains zero before the injected waterfront has reached a location; this observation suggests that the uncertainty component $P(A|C)$ and hence the integrated uncertainty $P(A|B,C)$ should remain zero until the locations are reached by the fluid front. Furthermore, the entropy at each location initially increases because of the influence of uncertainty at neighbouring locations. Evidently, the entropy in production response at the well locations declines as the flood sweeps through the bulk of the reservoir. These observations guided the development of the algorithm outlined below. All probabilities are assumed to be cumulative values:

1. Assume maximum prior uncertainty in particle count at all locations, i.e. $P(A)$ is a uniform distribution.
2. Compute the heterogeneity/convection related component of uncertainty $P(A|B)$ via indicator kriging.
3. At well locations: $P(A|C)_0$ is set to be a step function that increases its values from zero to one at the prescribed flow rate. An indicator flag is set to be one at the injector location. The indicator flags at the unswept locations are set to be zero. For all locations, $P(A|B,C)_0$ is initialised to be the same as $P(A|C)_0$.
4. For each time step, compute $P(A|C)_{n+1}$ at location i ($i = 1,\ldots, N$) via probability kriging, which is kriging of probability values at each threshold, using probabilities at locations that are within the range of influence as defined by the exponential covariance model representative of diffusion.
    - If the flag value of the conditioning location equals one (meaning the fluid front has reached the node), values of $P(A|B,C)_n$ from time step n are used for probability kriging; whereas if the flag value of the conditioning location equals zero (meaning it has not been reached by the fluid front), maximum uncertainty in particle count at the neighbouring locations should be used instead. This amounts to the uncertainty at i being influenced by the maximum uncertainty at node j ($j \neq i$) prior to the arrival of the front there. After the arrival of the front at j, the reduction in uncertainty at that location propagates to i. The indicator flag values at i are updated to be one if $P(A|C)_{n+1} \neq P(A|C)_0$.
    - The local uncertainty P(A|B) at data location is zero. Moreover, if the flood front has not yet approached the conditioning data location, the diffusive component remains equal to the initial value $P(A|C)_0$ and $P(A|B,C)_{n+1} = P(A|B,C)_0$. Skip Step 5.
5. Update $P(A|B,C)_{n+1}$ as in Eq. 6 and rescale $P(A|B,C)_{n+1}$ by its sum. This rescaling causes the calculated values to be legitimate cumulative probability values.
6. At production or injection well locations: $P(A|C)_{n+1} = P(A|C)_0$ and $P(A|B,C)_{n+1} = P(A|B,C)_0$

3.2 Case Study

Again, consider the same 2-D 50x50 reservoir. The hard data and variogram indicate the high probability of the presence of a high-permeability flow path located in the NE direction (azimuth angle = -45^o). Here is a summary of the available information

- A total of 32 locations with hard data: 10 located inside a potential high-permeability flow path area; another 10 located in a transition intermediate permeability area; and 12 values are distributed in the low permeability areas. The locations of the hard data are shown in Figure 2 (left).

- Variogram for high permeability thresholds: same as the one described in the case study in section 2. Variogram for low permeability thresholds: Spherical model, azimuth angle = 0^o, isotropic with range equals 900m.
- Range for the diffusive exponential variogram: 300m
- An injector is placed in the SE corner. A large injection rate is assumed.

Figure 2 (right) shows the entropy map obtained from indicator kriging. This represents the prior local uncertainty.

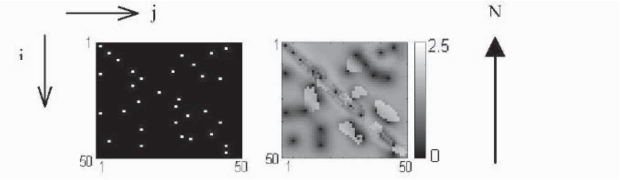


Figure 2. Left: location map of conditioning hard data. Right: entropy map from indicator kriging of the permeability field.

Figure 3 shows the entropy maps as a function of time obtained using the algorithm outlined above. Entropies are initially zero everywhere signifying that the fluid displacement has not commenced. As the front progresses, entropy begins to increase initially as the uncertainties in the surrounding nodes exert a strong influence on the uncertainty at a particular location; however, that uncertainty decreases as the fluid sweeps through the reservoir. The farther away the location is from the injection point, the longer is the time lag for the initial increase in entropy. Entropy remains low inside the high-permeability zone at all times because of the variogram structure and the conditioning influence of the data; water tends to sweep through the high permeability area before diffusing into the nearby low permeability matrix. An interesting point to be noted is that entropy remains relatively higher at regions that are close to the low-permeability hard data. This suggests that uncertainties are higher even at locations that are close to hard data, if the hard data is in a region of low permeability and is away from the main fluid flow paths. It can be seen that the entropy map in Figure 3 exhibits significantly more variability than the kriged entropy map (Figure 2). In a sense, the entropy updating process using the principle of superposition re-introduces covariance reproduction to the extent that is relevant from a flow perspective. Higher uncertainty is consistently indicated in the transition from the high to low permeability zone, and this would also be observed in the local and joint uncertainty distributions obtained using kriging and stochastic simulation.

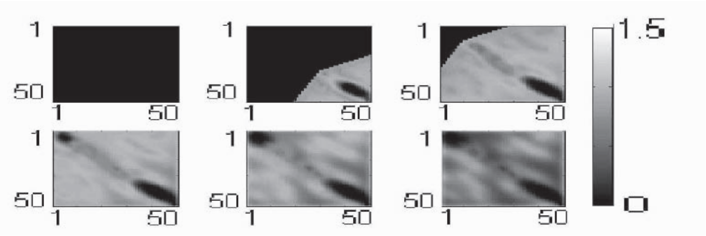


Figure 3. Entropy maps at various snapshots of time. From top left to bottom right: *t* = 0, *t* = 10, *t* = 20, *t* = 30, *t* = 50, *t* = 100.

As a comparison, fifty realizations based on the given information were generated using sequential indicator simulation. Each of the realizations was then subjected to flow simulations. Each flow simulation case has an injector at node (50,50), injecting water at a rate of 250rm^3/day; a producer flowing at 250rm^3/day is located at one of 6 selected locations whose locations are marked in Figure 4. Figure 4 (left) shows the entropy of water cuts at various producer locations as a function of time. In general, entropy fluctuates slightly at the initial times due to connate water production. Once the injected waterfront has reached the producer location, entropy starts increasing as water cut increases. The entropy continues to increase to a maximum, and then a gradual reduction is observed as time progresses. As seen in Figure 4 (left), locations that are close to the injector, such as (48,48), experience the influences of the injected water much sooner than locations that are far away, such as (3,3). It is noted that the entropies are generally high for low-permeability locations, and it also takes longer for the entropy to decline at those locations. The rate of decline also depends on distance from injection point. The farther away from injection point, the more gradual is the decline in entropy. Nonetheless, the entropy peaks appear to be of similar magnitude for all locations considered. Other useful statistics to consider would be uncertainties in breakthrough times and the water cuts at some time after breakthrough.

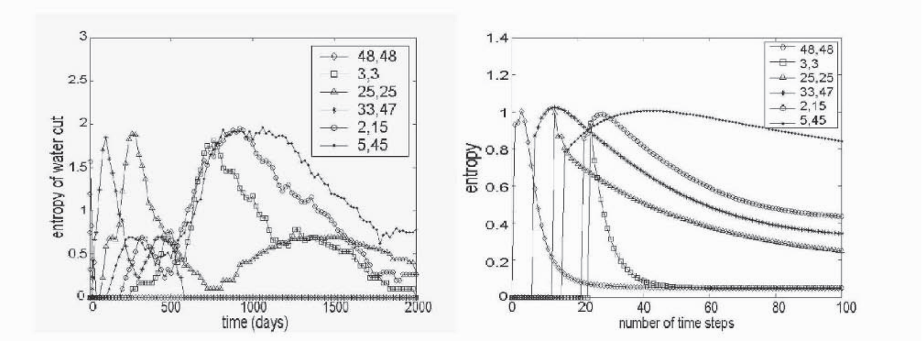


Figure 4. Entropy profile at six selected locations (Left: flow simulations; right: our direct uncertainty assessment technique)

Figure 4 (right) shows the entropy profile as a function of time for the same six selected locations, obtained using our direct assessment technique. Similar to the results obtained from flow simulations over 50 realizations, entropy peaks at all locations are approximately the same. The entropy peaks inside the high-permeability zone are slightly lower; entropies in low-permeability regions reduce much more gradually than those in the high-permeability regions. The order in which breakthroughs and entropy peaks at various locations occur is also the same as in flow simulation results. Entropy at location (5,45) remains high at all times, and a similar trend can be observed in results from flow simulations as well. The magnitude of the entropy value at a location immediately after the arrival of the flood front is equivalent to the entropy in breakthrough time. The uncertainty in breakthrough in the high permeability zone decreases slightly as the distance from the injection point increases. This might be due to the fact that for locations that are close to the injection point, any fluctuations in flow properties at the injection point would manifest itself immediately on the flow uncertainty at locations that are close to the injection point. It can be seen that a small amount of residual entropy or uncertainty exists for all locations even after a large number of time steps. Unlike results from flow simulations, the initial increase in entropy peak predicted by our method occurs abruptly. This is a consequence of the indicator flag being turned on abruptly when a front reaches a location. In the finite difference simulation, there is a smearing of the flood front as it approaches a location.

One of the major application areas of uncertainty assessment procedures such as that described in this paper is well placement optimization. Important variables to consider include uncertainty in breakthrough times and uncertainty in water cut after breakthrough. Our stochastic flow simulation provides a proxy breakthrough time indicator, which is defined as the instant when the flag at a location changes from zero to one, i.e. the moment a location is reached by the fluid front. In order to gauge the uncertainty in breakthrough, we can look at the entropy value immediately following breakthrough at each location. As explained, the initial increase in entropy occurs because of the propagation of uncertainty from surrounding nodes. Based on these observations, if we are to decide the most optimal locations for a producer well, in order to achieve the maximum efficiency, we would like to place the well at a location where water breakthrough occurs late and the uncertainty associated with that breakthrough is low. As a result, a location with long breakthrough time, and yet low entropy peak is preferred. Details about the applications of the proposed approach to practical reservoir management can be found in Leung (2004).

4 Conclusions & Recommendations

We demonstrated a fast transfer function methodology where the convective and diffusive components of transport are decoupled. As an application of this simple flow model, we proposed a new technique for direct uncertainty assessment via a single stochastic flow simulation that combines local geologic uncertainty with a simple flow transfer model. Unlike the traditional Monte Carlo approach, where constructing the uncertainty distribution location by location can be tedious, the full uncertainty distribution is available at every location as a function of time.

This page intentionally blank

References

Bu, T. and Damsleth, E. (1996). Errors and uncertainties in reservoir performance predictions. *SPE Formation Evaluation*, September, 194-200.

Crank. J. (1975). *The Mathematics of Diffusion*. Oxford: Oxford University Press.

Journel, A.G. (1994). Modelling uncertainty: some conceptual thoughts. In R. Dimifrakopoulos (Ed.). *Geostatistics for the next century* (pp. 30-43). Dordrecht: Kluwer Academic Press.

Leung, J. Y. (2004). *A new analog for numerical flow simulations. Application to direct assessment of reservoir flow performance uncertainties*. Master's Thesis, University of Texas at Austin.

Lia, O., Omre, H., Tjelmeland, H., Holden, L., and Egeland, T. (1997). Uncertainties in reservoir production forecasts. *AAPG Bulletin*, 81(5), 775-802.

Sabelfeld, K. and Kolyukhin, D. (2003). Stochastic Eulerian model for the flow simulation in porous media. *Monte Carlo Methods and Applications*, 9(3), 271-290.

Shannon, C.E. (1948). A mathematical theory of communication. *The Bell Systems Technical Journal*, 27, 379-423.

Sudaryanto, B. and Yortsos, Y.C. (2000). Optimization of fluid front dynamics in porous media using rate control. I. Equal mobility fluids. *Physics of Fluids*, 12(7), 1656-1670.

Sudaryanto, B. and Yortsos, Y.C. (2001). Optimization of displacements in porous media using rate control. Paper SPE 71509 presented at the *SPE Annual Technical Conference and Exhibition*. New Orleans, Louisiana, U.S.A., Sept. 30 – Oct. 3, 2001.

PRESERVATION OF MULTIPLE POINT STRUCTURE WHEN CONDITIONING BY KRIGING

WEISHAN REN, LUCIANE CUNHA AND CLAYTON V. DEUTSCH
Department of Civil & Environmental Engineering, 220 CEB,
University of Alberta, Canada, T6G 2G7

Abstract. The idea of conditioning by kriging is well known in theory and practice. It has been used for conditioning the realizations from unconditional simulation techniques such as the moving average and the turning bands simulation approaches. The basis of the conditioning by kriging approach is to use the same variogram for both the unconditional simulation and the kriging. In this paper, the focus is on using kriging for conditioning of more complex unconditional simulations. Unconditional simulated realizations with multiple point structure are generated for posterior conditioning. Two sets of data are used for kriging. After conditioning, the simulated values at data locations are the real data values, so the local data is honored. Beyond the range of correlation, the simulated values are the unconditional simulated values, which mean that the multiple point structure can be preserved.

The results obtained in this work show that conditioning by kriging is a simple, easy and reliable way to account for data with complex multiple point structures. Both continuous and categorical variables are used to show the performance of the conditioning by kriging approach. Moreover, kriging with different sets of data demonstrates that the multiple point structures are well preserved after conditioning.

1 Introduction

Conditioning by kriging is not a new concept in geostatistics. It has been used for conditioning the moving average and the turning bands simulations (Journel and Huijbregts, 1978). However, because the sequential simulation is quite feasible and widely accepted, the conditioning by kriging has been little used in practice.

When working on reservoir characterizations, we often need to deal with complex features. Two-point statistical simulations are not good at preserving these complex features. So multiple-point statistic is usually needed. However, multiple-point statistic is very complicated and always takes a very long time. Object-base simulation can also be used to model the complex features. But conducting a conditional simulation with complex features takes much longer time than an unconditional simulation. If carrying out an unconditional simulation to capture these complex features, then apply the conditioning by kriging to obtain a conditional simulation, the conditional simulation with complex features can be performed in an easy and fast manner. In this paper, the

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 643–652.

focus is on the conditioning by kriging. The unconditional simulation realizations are generated by using the *fluvsim* (Deutsch and Tran, 2001) and contain channels and overbank with non-linear features. The conditioning by kriging is conducted to show how these multiple point structures can be preserved. The performance of conditioning by kriging will be shown in both continuous and categorical variables cases.

2 Theory of Conditioning by Kriging

Suppose there are N real data, and M data need to be simulated. The conditioning by kriging consists of the following successive steps:

1. Carry out an unconditional simulation to obtain the unconditional simulated values $Z_{uc}(x)$.
2. Carry out kriging using the N conditioning data to obtain $Z_{kr}(x)$.
3. Carry out kriging using the unconditional simulated values at these N data locations to obtain $Z_{kr-u}(x)$.
4. Calculate the conditional simulation values for each block:

$$Z_{cs}(x) = Z_{uc}(x) - [Z_{kr-u}(x) - Z_{kr}(x)]$$

This equation implies that at each real data location, the unconditional simulated value is taken out, and the conditioning datum is put in. Near the location, the kriging linear estimators smooth the change between the real data and the unconditional simulated values outside the range of kriged values. Therefore, after the conditioning, the conditional simulated values at these N data locations will exactly be the real data values. Beyond the range of correction, the conditional simulated values will be the unconditional simulated values. These steps need to be carried out in the Gaussian environment.

The two kriging (steps 2 and 3) can be combined to perform only one kriging using N data differences between the real data and the unconditional simulated values. Then the conditioning can be simply expressed as:

$$Z_{cs}(x) = Z_{uc}(x) + D_{kr}(x)$$

where $D_{kr}(x)$ is the data of the kriging using the residual of the real data values subtracting the unconditional simulated values.

It can be interpreted as that the conditioning by kriging is adding correction areas to unconditional simulations based on the differences between real data and unconditional simulated values. This can be easily seen in an example (Figure 1). This example is constructed using categorical variables. The unconditional simulation realizations and the string of conditioning data are shown together in the left image. The result of conditioning by kriging is shown in the right image. The channel is in dark grey, and the overbank is in light grey. For the places in overbank where the conditioning data shows there should be a channel, a certain size of channel is added by conditioning. In the example, a channel is added beside the original channel so that the result looks like the original channel is dilated. In the contrary, for the places where the conditioning data shows there should be no channel, erosion takes place.

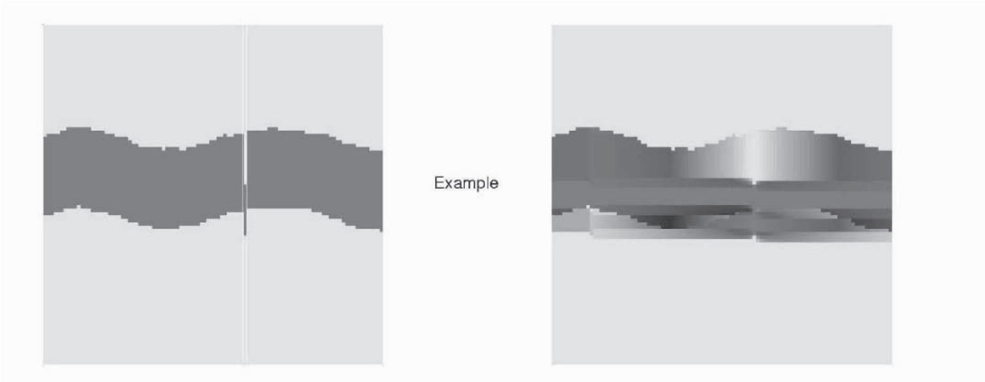


Figure 1. Example of conditioning by kriging.

Continuous and categorical variables are used to show the performance of the conditioning by kriging approach. The real data can be obtained from continued cores in a single well or from different wells. Therefore, for each variable, the conditioning data use a string of data and scattered data. They are extracted from an unconditional simulation.

3 Continuous Variable Cases

3.1 CONDITIONING WITH A STRING OF DATA

A string of data used for the conditioning is shown in Figure 2. The real data actually are only one pixel wide. In order to show them clearly in a plot, the string of data is extended to five pixels. The data domain is 100 by 100.

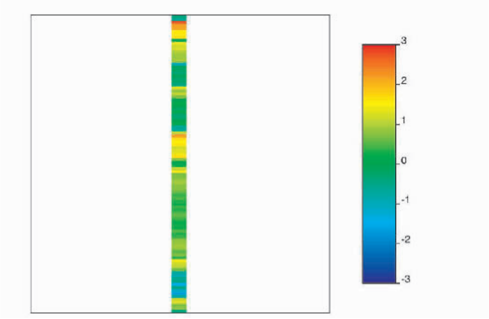


Figure 2. A 1-D string of conditioning data expanded to five pixels. The real data is one pixel wide at index 50 in a 100 x 100 pixel image. The normal scale is used throughout.

The unconditional simulation was implemented using the *fluvsim* to generate realizations with channels and overbank deposits. Within the channels and overbank deposits, a separated *sgsim* (Deutsch and Journel, 1998) was implemented to generate the realizations for each deposit. The four unconditional simulation realizations were transformed into Gaussian space and are shown in Figure 3. These curvilinear channels are laying out in the vertical direction. The normal scores of these unconditional simulation realizations were used to calculate variograms. The *varfit* (Larrondo *et. al.,* 2003) was used to model the variograms, and the variogram models were used in the kriging. The models of variograms in the vertical and horizontal directions of the first unconditional simulation realization (the top left image in Figure 3) are shown in Figure 4. The residual data was calculated by subtracting the unconditional simulated values from the conditioning data values. Therefore, using these residual data, only one kriging was implemented. The kriged values were added to the unconditional simulated values to calculate the conditional simulation values.

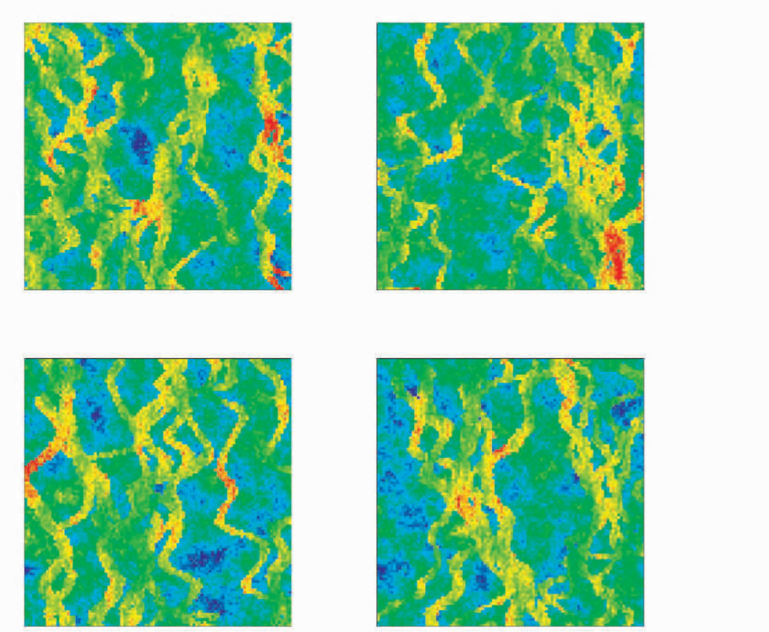


Figure 3. Four unconditional realizations generated by *fluvsim* and then separate *sgsim* runs within the channel and overbank deposits.

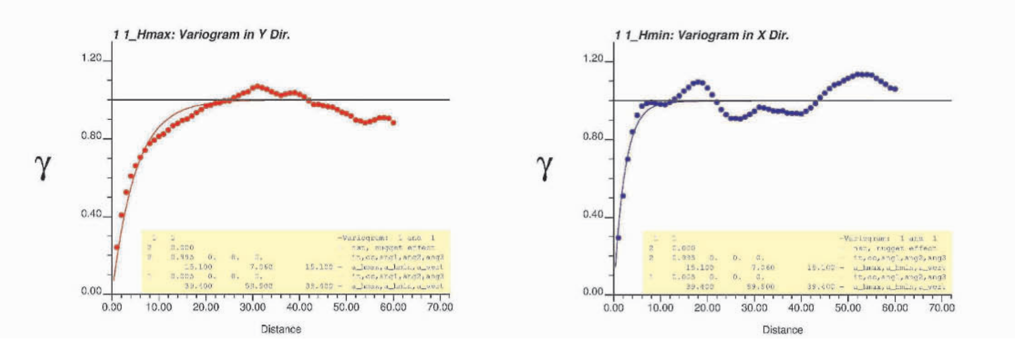


Figure 4. Variograms of the first unconditional realization in two main directions.

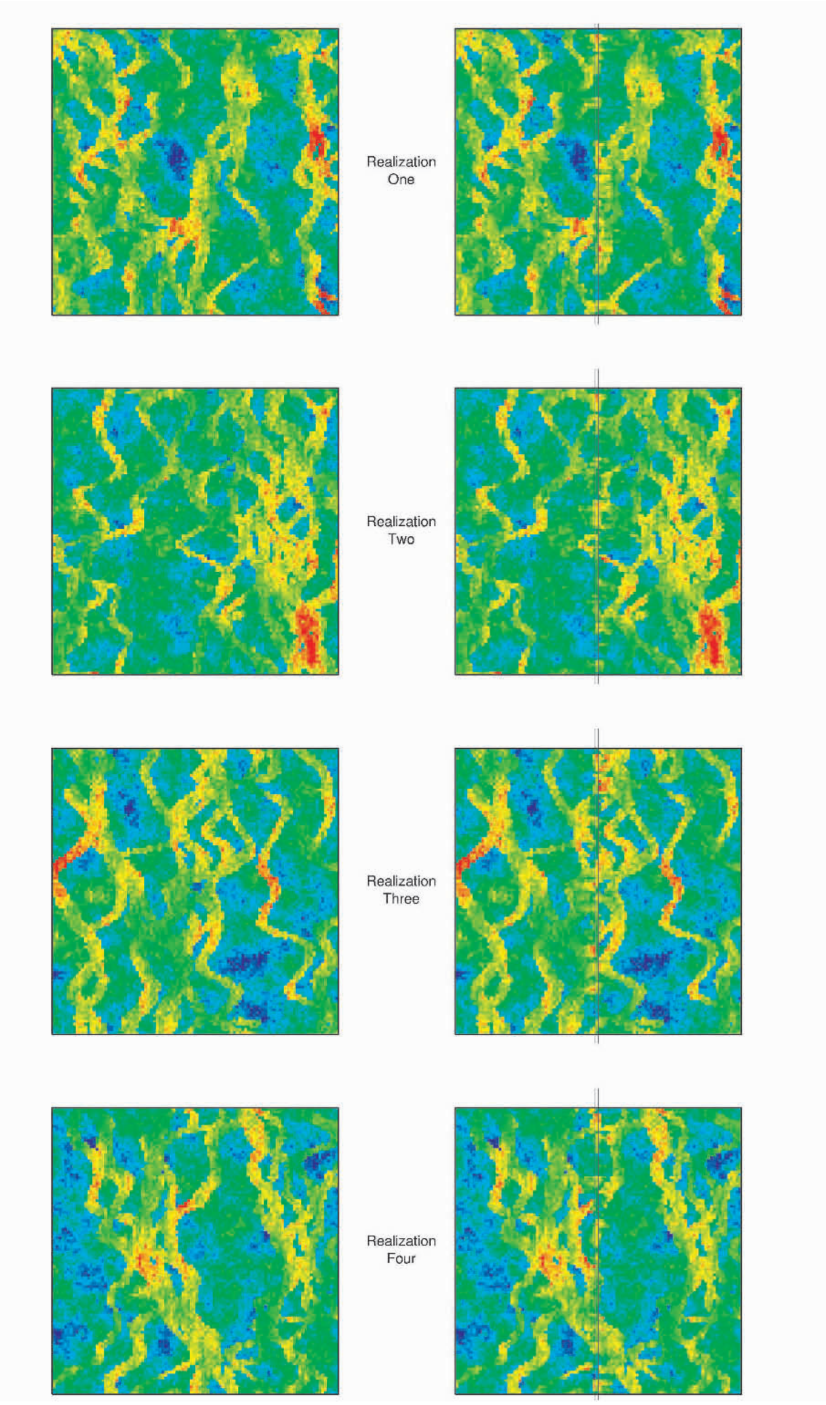


Figure 5. The left column is four unconditional simulation realizations, and the right column is these realizations conditioned by kriging. The central data location matches exactly the conditioning data shown on Figure 2.

The conditional simulation realizations are shown in the right column of Figure 5. For comparison, the four unconditional simulation realizations are plotted in the left column of Figure 5. It can be seen that near the center line of the realizations, there are some difference created by the conditioning. Beyond that, they are exactly same. Apparently, the multiple-point structures are well preserved.

3.2 CONDITIONING WITH SCATTERED DATA

The scattered data used for the conditioning by kriging are shown in Figure 6. The data are taken at ix/iy grid node indices of 10, 30, 50, 70 and 90.

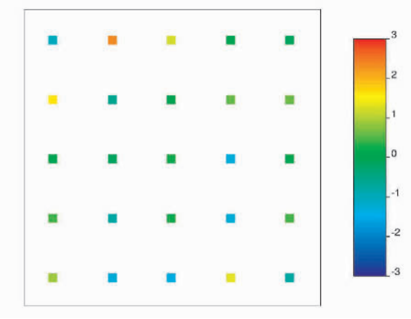


Figure 6. Scattered conditioning data: the data are taken at ix/iy grid node indices of 10, 30, 50, 70, and 90.

The same approach has been applied for conditioning the four unconditional simulation realizations (Figure 3) with the 25 scattered data. Both the conditional simulation results and the unconditional simulation results are shown in Figure 7. The conditioning data are honored and the multiple-point structures are well preserved.

4 Categorical Variable Cases

4.1 CONDITIONING WITH A STRING OF DATA

Similarly to the continuous variable case, a string of data was used for conditioning (Figure 8). The data are categorical values of 1 and 0. The channel categorical value is 1 and shown in dark grey. The overbank categorical value is 0 and shown in light grey. The real data is only one pixel wide at index 50 in a 100 by 100 pixel image, and it is also expanded to five pixels to show them better. Two unconditional simulation realizations (the left column in Figure 9 or 11) were generated with *fluvsim*. The data are also categorical values of 1 and 0. These curvilinear channels are laying out in the vertical direction. The despike was used to change the same categorical values into slightly different values so that the 1 : 1 normal score transformation could be achieved. The categorical values of the unconditional simulated data and the conditioning data were despiked, and transformed into normal scores. The variograms were calculated from the unconditional realizations and modeled by the *varfit* program. These models were used in the kriging.

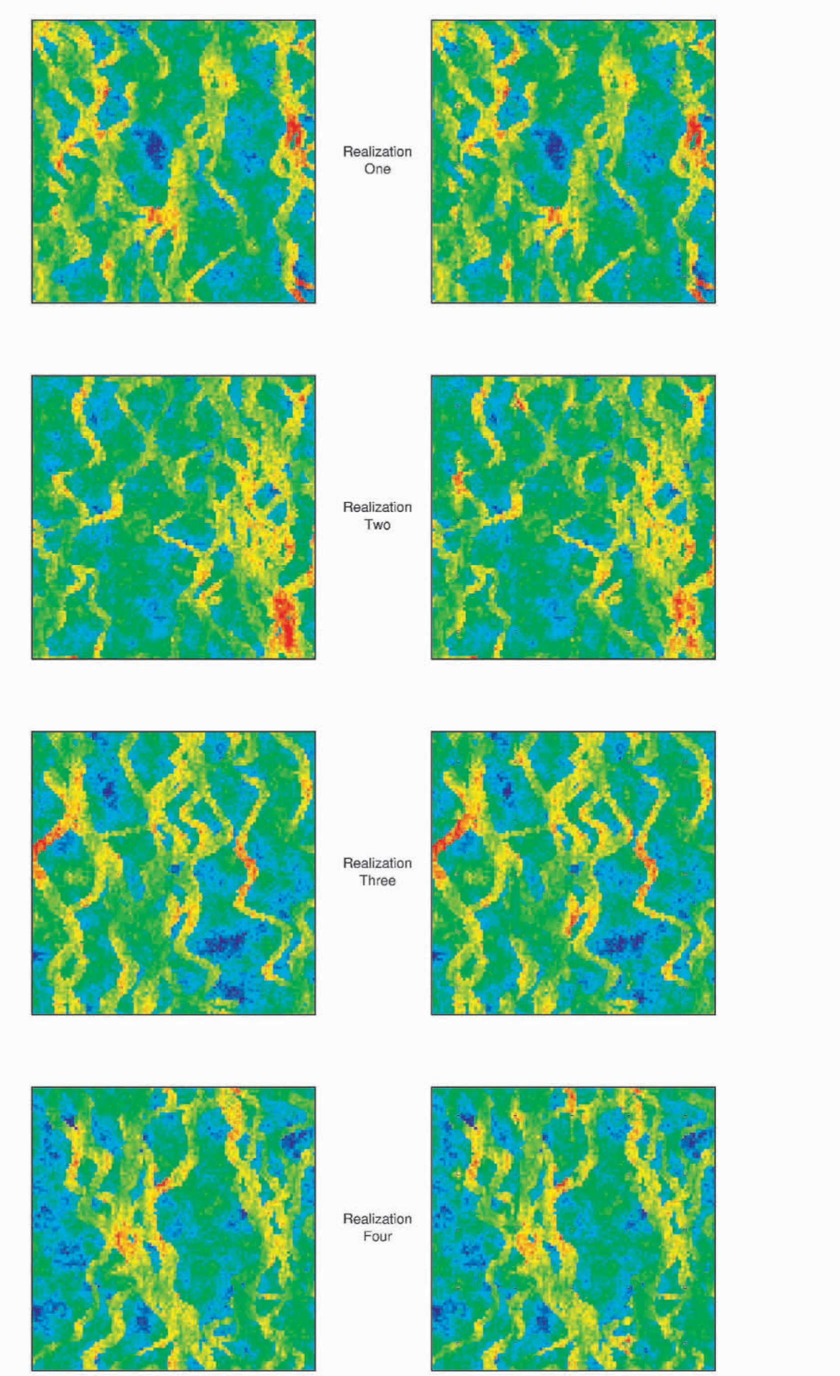


Figure 7. The left column is four unconditional simulation realizations, and the right column is these realizations conditioned by kriging to 25 scattered data. They match the conditioning data exactly and the non-linear structure is preserved quite closely.

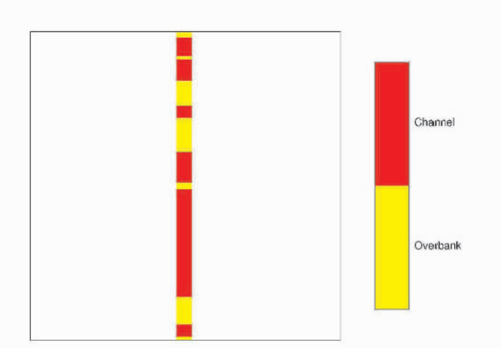


Figure 8. A string of conditioning data for the categorical variable case. The data is expanded to five pixels.

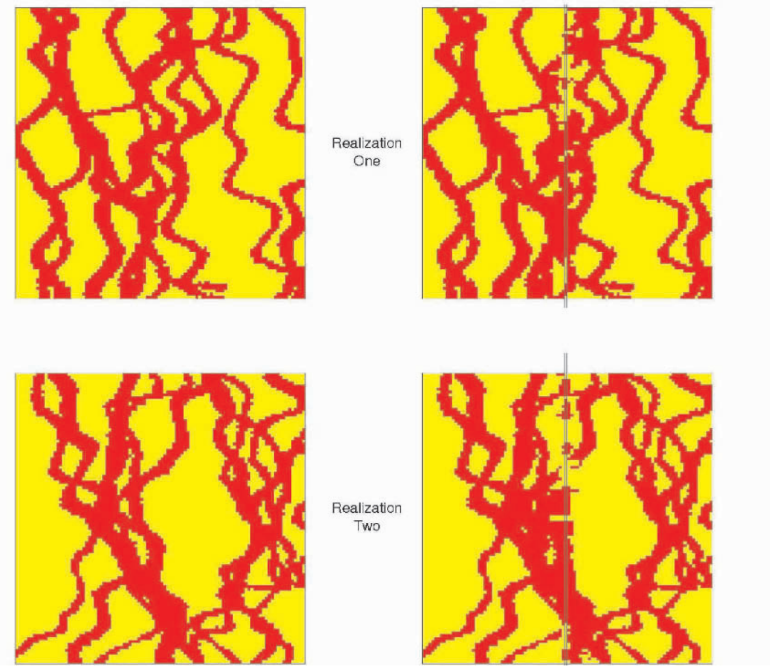


Figure 9. Two unconditional simulation realizations are on the left, and the realizations conditioned by kriging to the string of data are on the right. These conditioned models match the conditioning data.

The residual data were calculated in Gaussian space. After kriging with the residual data, the kriged results were added to the unconditional simulated data to get the conditional simulated data. These values were normal scores but not categorical values. So they were truncated into categorical values of 1 and 0.

The conditional simulation realizations are shown in the right column in Figure 9. The models on the right match the conditioning data. There are some artifacts when large changes occur, but apparently, the non-linear structures are well preserved.

4.2 CONDITIONING WITH SCATTERED DATA

The scattered data used for the conditioning by kriging are shown in Figure 10. The data are taken at ix/iy grid node indices of 10, 30, 50, 70 and 90. The same approach has been applied for conditioning the unconditional simulation realizations. Both the unconditional and the conditional simulation results are shown in Figure 11. Some artifacts appear when facies changes. But certainly the multiple-point structures are well preserved.

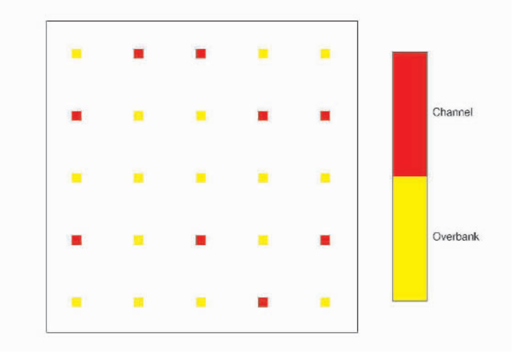


Figure 10. 25 scattered conditioning data for the categorical variable case. The data are taken at grid node indices of 10, 30, 50, 70 and 90.

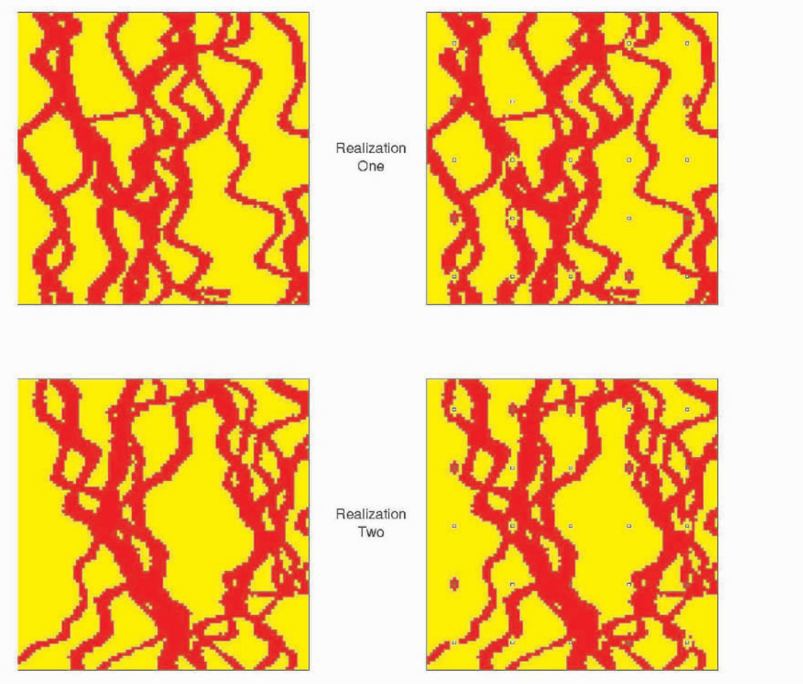


Figure 11. Two unconditional simulation realizations are on the left, and the realizations conditioned by kriging to the scattered data are on the right. These conditioned models match the conditioning data.

5 Conclusions

Conditioning by kriging is correct when dealing with unconditional realizations based on the simple kriging principle. The conditional realizations respect the data and have no artifacts of the conditioning data. Using kriging to condition realizations that are generated by more complex simulation algorithms has been demonstrated with limited success. The shape/structure of the changes near the conditioning data respects mostly the variogram. The greater the change required, the greater the influence of the variogram and the poorer the complex structure is preserved. The algorithm would work well when the changes are minimal (such as a near solution with annealing) or when the features are reasonably captured by the variogram.

References

Deutsch, C.V. and Journel, A.G., *GSLIB: Geostatistical Software Library and User's Guide*, 2nd Edition, Oxford University Press, 1998.

Deutsch, C.V. and Tran, T.T., FLUVSIM: a program for object-based stochastic modelling of fluvial depositional systems, Computers & Geosciences, vol. 28, 2002, p. 525-535.

Larrondo, P.F., Neufeld, C.T. and Deutsch, C.V., VARFIT: A program for Semi-Automatic Variogram Modelling. In *Center for Computational Geostatistics Report Five,* University of Alberta, Edmonton, Alberta, 2003.

Journel, A.G. and Huijbregts, C.J., *Mining Geostatistics*, Academic Press, 1978.

STOCHASTIC MODELING OF THE RHINE-MEUSE DELTA USING MULTIPLE-POINT GEOSTATISTICS

AMISHA MAHARAJA
Department of Petroleum Engineering, Stanford University, Stanford, CA 94305

Abstract. An extensive data set on the Rhine-Meuse delta in the Netherlands is available. It consists of approximately 200,000 boreholes and fully reconstructed channel systems for the past 10,000 years. In a study area the reconstructed channel systems are treated as ground truth for checking the simulation results from the multiple-point stochastic modeling algorithm *snesim*. Two generations of channel systems with distinct morphological characteristics can be identified in the study area. These are modeled jointly, then separately. Joint simulation gives poor results because the individual characteristics of each channel system are lost when taken together. When simulated separately, the attributes of the two different channels systems are better reproduced, the simulations are then combined by cookie-cut.

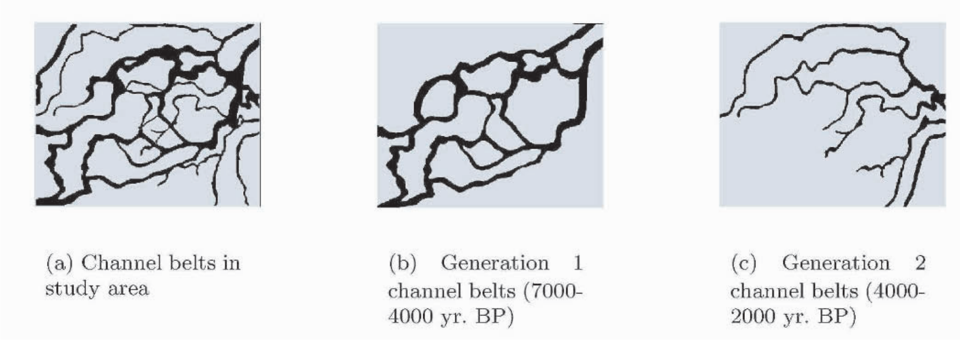
Key words: Multiple-point statistics, Hierarchical simulation, Delta

1 Introduction

The subaerial part of the Rhine-Meuse delta during the Holocene has been extensively studied by researchers of Utrecht University in the Netherlands (Berendsen and Stouthamer, 2001). An area of 5700 sq. km. was extensively drilled and channel belts of the ancient river systems have been reconstructed. Over the years some 200,000 lithologic boreholes have been sampled and described. What makes this data set unique is that the paleogeographic reconstruction has been incorporated into a GIS data base so that channel belts existing at any given time in the past 10,000 years can be easily retrieved (Cohen, 2003). Because of this feature, the data set lends itself well to testing stochastic modeling algorithms.

We selected an area of 11.4 × 9.36 sq. km. (600 × 500 grid nodes) within the delta area (Figure 1(a)) to test the multiple-point simulation algorithm *snesim* (Strebelle, 2000, 2002). The maximum thickness of the Holocene deposit in the area is approximately 13 meters, hence the vertical stacking patterns are uninteresting. Instead, we focus on high resolution modeling of the 2D channel patterns which have been worked out in detail by the Utrecht researchers. The channel belts in this area can be separated into generation 1 (Figure 1(b)) and generation 2 (Figure 1(c)).

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 653–659.



(a) Channel belts in study area

(b) Generation 1 channel belts (7000-4000 yr. BP)

(c) Generation 2 channel belts (4000-2000 yr. BP)

Figure 1. Channel belts of the Rhine-Meuse delta in the selected study area

One reason for modeling these two generations separately is that they are significantly different in their channel patterns, orientation, and channel widths. The generation 1 channel system is an anastomosing channel system oriented at roughly 45^o from North in which the channel belts join and re-join forming networks of interconnected channel belts. On the contrary, the channel belts of generation 2 have an open pattern and they fan out. The width of generation 1 channel belts is in 100s of meters while that of generation 2 is in 10s of meters. Moreover, the NTG of generation 1 is 20 percent, which is double the NTG of generation 2. The following sections describe the conditional simulation of the two generations jointly and separately.

2 Joint simulation

For the joint simulation of both generations of channel systems, the reference GIS image with a net-to-gross (NTG) of 28 percent is used as a training image (Ti), see Figure 1(a). Fifty well data with a NTG of 28 percent are used as hard data. A high resolution stratigraphic modeling would generally not be possible without seismic data in actual hydrocarbon reservoirs. Since no seismic data is available, a local probability map for occurence of channel facies is generated by taking a moving window average of the reference image with a 50 x 50 moving window (Figure 2). A target NTG of 29 percent is enforced by the servo-system provided in the program *snesim*. Simulations conditional to the hard well data and the soft probability data are generated. For details about how the various data are honored in the snesim algorithm refer to Zhang (2003).

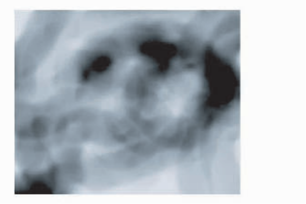


Figure 2. Channel probability map for generation 1

3 Separate simulation

For simulation of generation 1, a specific Ti (Figure 3) is drawn using Geobody-Painter software (Frank, 2003). This Ti captures the anastomosing pattern of generation 1 channel system, however, the channel belts are not oriented at 45^o from North. This rotation information, typically obtained from seismic data interpretation, is provided directly as an input to the simulation algorithm. Since generation 1 channel belts do not occur outside a certain area (Figure 1(b)), a no-simulation region is defined so that no channels are simulated there. Generation 1 well data with a NTG of 16 percent are provided and a target NTG of 20 percent is enforced with the program servo-system. Simulations conditioned to the hard well data, the region information and the global rotation information are then generated.

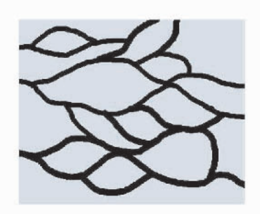
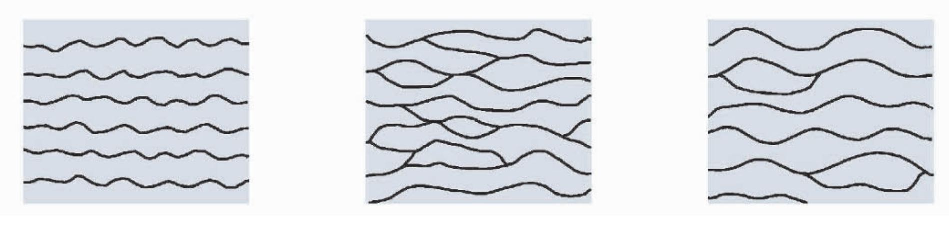


Figure 3. Training image for Generation 1

Training images shown in Figure 4 are drawn for simulation of the generation 2 channel system. The Ti in Figure 4(a) has straight channels with short-scale undulations without any loop. The Tis in Figure 4(b) and 4(c) have looping channels with large-scale undulations. A local probability map for occurence of channel facies is generated by running over the reference image a 50 x 50 moving window average (Figure 5(a)). A locally varying angle map (Figure 5(b)), derived from the soft data, is supplied to achieve the faning pattern. Generation 2 well data with a NTG of 22 percent are used as hard data and a target NTG of 10 percent is enforced through the program servo-system. Simulations conditioned to the hard well data, the soft probability map and the local rotation map are then generated.

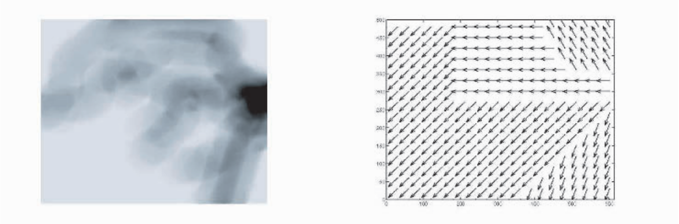


(a) Training image 1

(b) Training image 2

(c) Training image 3

Figure 4. Training images for Generation 2



(a) Soft probability map

(b) Angle map

Figure 5. Generation 2 data

4 Discussion of simulation results

When taken together, generation 1 and generation 2 form a mesh of wide and narrow channel belts which completely obscures their individual patterns (Figure 1(a)). The modeling algorithm cannot "see" the individual components and averages out the differences in channel widths. Consequently, the resulting simulation, although exactly data-conditioned, bears little resemblance to either generation 1 or generation 2 (Figure 6). The NTG of the realization is 28.4 percent, which is close to the target 29 percent.

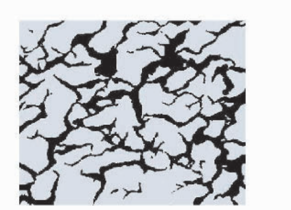
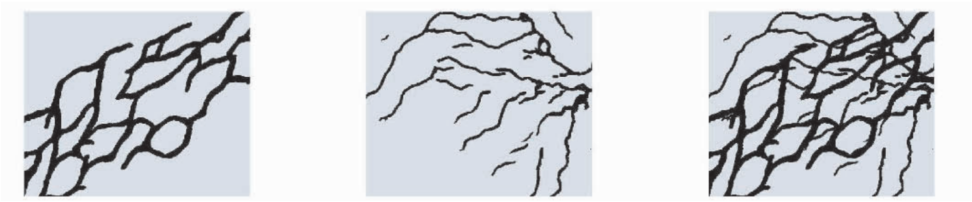


Figure 6. Joint simulation of generation 1 and 2 channel belts

Figure 7(a) and 7(b) shows realizations of generation 1 and 2 individually and Figure 7(c) shows the result after generation 2 is cookie-cut onto generation 1. A different set of Tis, angle, and soft data were used for generation 1 and 2, which made it possible to preserve their individual characteristics. The NTG of generation 1 realization is 20.2 percent, while that of generation 2 is 10.5 percent. The NTG after cookie-cut is 28 percent, which is close to the target 29 percent.



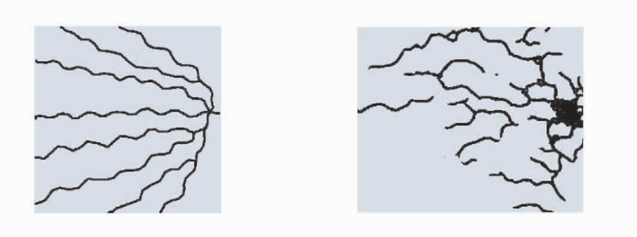
(a) Generation 1 simulation

(b) Generation 2 simulation

(c) Generation 2 cookie-cut onto generation 1

Figure 7. Separate simulation of generation 1 and 2 channel belts

For the simulation of generation 1 channel belts, it was essential to define the region of no-simulation in addition to providing a Ti with correct channel width and looping pattern and the global rotation information. Simulation of generation 2 channel belts required the additional constraint of local angle map (Figure 5(b)). The soft probability cube guided the channel density to give a better match with the reference image. Better result is obtained by using a stationary Ti (Figure 4(a)) supplemented with an angle map than using a non-stationary Ti that includes the angle information (Figure 8).

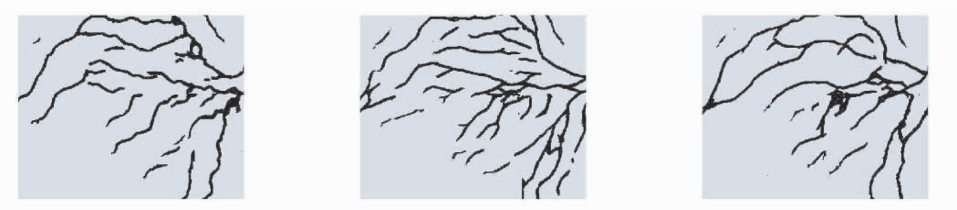


(a) Fan Ti (600 x 600)

(b) Realization

Figure 8. Simulation of Generation 2 using fan Ti

The simulations of generation 2 corresponding to the three different Tis in Figure 4 are shown in Figure 9. The simulation (Figure 9(a)) using the Ti of Figure 4(a) best resembles the reference generation 2 channel pattern. This is because generation 2 channels are mostly free of loops and have short-scale undulations, which is reflected in the Ti of Figure 4(a). However, connectivity is poor because the NTG is low and the channels in the Ti are not connected. Better connectivity is obtained using the Ti of Figure 4(b) because there the channels display connections, see Figure 9(b). The channels in the Ti of Figure 4(c) have fewer inter-channel connections, hence the connectivity in the corresponding simulation is poorer, see Figure 9(c).



(a) Simulation using Ti 1

(b) Simulation using Ti 2

(c) Simulation using Ti 3

Figure 9. Simulation of Generation 2 using different Tis

In general, it is important to specify the correct channel widths to get correct channel density. If the channels in the Ti are wider than in reality, then the target NTG can be attained with fewer channels resulting in a lower density of simulated channels. Conversely, if the training channels are thinner than in reality, then a greater number of channels will be simulated in order to match the target NTG.

In the case of generation 2, channel probability map alone is not sufficient to impose the correct angle when simulating with a stationary Ti (Figure 4(a)). This is because in the Ti, there is no data event in which the channels are oriented at an angle other than 90 degrees. The textitsnesim program does not extract the angle information from the soft probability map. The user has to generate the angle map and provide it explicitly to *snesim* as an additional constraint, as shown in case of generation 2 simulation.

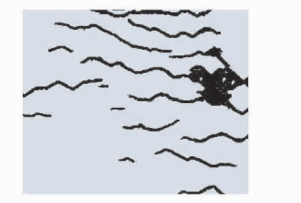


Figure 10. Simulation of generation 2 channel belts without local angle data (613 x 500)

5 Conclusions

This application of multiple-point stochastic simulation on the Rhine-Meuse delta brings out several important point for modeling such deposits using the /textit-snesim algorithm.

- **Need to separate depositional sequences**: If geological interpretation indicates the presence of distinct channel systems, these systems should be simulated separately to preserve their specific structures. A hierarchy of simulation may also be applied within the same system when multiple facies with different shapes and sizes are present (Maharaja, 2003).
- **Importance of accurate training images**: It is important to provide Tis with the correct channel characteristics such as (in this case study) width, roundness and size of the channel loops, otherwise the Ti information may conflict with the seismic and well data. Moreover, channel density will be poorly reproduced if the Ti channels are too wide or too narrow.
- **Importance of local angle data**: Soft probability data alone cannot impose the angle information because the textitsnesim program does not extract the angle information from the channel probability map. Such locally varying angle information needs to be provided explicitly by the user.
- **Importance of stationary Ti**: With *snesim*, it is better to use a stationary Ti. Non-stationarity can be imparted by using additional local information such as angle maps and local NTG proportion maps. An area that is deemed non-stationary should be divided into regions which are then simulated separately with corresponding stationary Tis. To avoid discontinuity between simulated regions, these could be simulated sequentially using overlapping data templates. Pattern-based algorithms such as *simpat* and *filtersim* do

This page intentionally blank

not average out features of the Ti unlike *snesim*, hence they do not require stationary Tis.

Acknowledgements

We would like to acknowledge the industry sponsors of the Stanford Center for Reservoir Forecasting at Stanford University for supporting this research. Special thanks to Professor Berendsen of Utrecht University for providing the data set.

References

Arpat, B. and Caers, J., *A Multiple-scale, Pattern-based Approach to Sequential Simulation*, Proceedings of the Seventh International Geostatistics Congress, Banff, Canada, 2004.

Berendsen, H.J.A. and Stouthamer, E., *Paleogeographic development of the Rhine-Meuse delta, The Netherlands*, Koninklijke Van Gorcum - Assen, The Netherlands, 2001.

Cohen, K., *Differential Subsidence Within a Coastal Prism*, PhD thesis, Utrecht University, Utrecht, The Netherlands, 2003.

Frank, T., *GeobodyPainter Plug-in, version 1.0.0*, Gocad Research Group in Nancy, France, 2003.

Maharaja, A., *Hierarchical Simulation of Multiple Facies Reservoirs Using Multiple-point Geostatistics*, MS thesis, Stanford University, Stanford, CA, 2004.

Strebelle, S., *Sequential Simulation Drawing Structures from Training Images*, PhD thesis, Stanford University, Stanford, CA, 2000.

Strebelle, S., *Conditional Simulation of Complex Geological Structures Using Multiple-point Statistics*, Mathematical Geology, vol. 34, no. 7, 2002, p. 1161-68.

Zhang, T., *Merging Prior Structural Interpretation and Local Data: The Bayes Updating of Multiple-point Statistics*, SCRF Report 16, Stanford University, Stanford, CA, 2003.

Zhang, T., *Sequential Conditional Simulation Using Classification of Local Training Patterns*, Proceedings of the Seventh International Geostatistics Congress, Banff, Canada, 2004.

STOCHASTIC SIMULATION OF UNDISCOVERED PETROLEUM ACCUMULATIONS

ZHUOHENG CHEN[1], KIRK G. OSADETZ[1], HAIYU GAO[2]
[1]Geological survey of Canada, 3303-33rd Street NW, Calgary, AB, T2L 2A7
[2]Schlumberger Information Solutions, Calgary Technology Center, 600 322-11 Avenue SW. Calgary, T2R 0C5, Canada

Abstract. Geographic characteristics of undiscovered petroleum accumulations are important both to better natural resource management and improved exploration efficiency. Stochastic simulation is a useful tool that reveals uncertainties in petroleum exploration and exploitation applications. The lack of information regarding the locations of undiscovered petroleum accumulations presents a major difficulty to the application of this technique to petroleum resource assessment. In order to facilitate the locations of undiscovered petroleum accumulations, we propose a model-enhanced simulation approach that uses a geological model, in either the form of geological favorability or probability of petroleum occurrence derived from available geological and geophysical observations. The proposed approach employs a Fourier transform algorithm in the conditional simulation because it permits the spatial correlation-specific and location-specific features from different data sources to be studied separately and integrated in the frequency domain subsequently. This approach is illustrated by the analysis of the Rainbow petroleum play in the Western Canada Sedimentary Basin. The proposed approach produces a resource map showing the possible size of undiscovered petroleum accumulations with geographic locations. A comparison with the results from a traditional conditional simulation indicates that the proposed approach produces maps with improved features and predictions validated by the test data set.

1 Introduction

During the past three decades, petroleum resource assessment method development has focused primarily on assessing the aggregated properties of oil and gas resources, such as the total potential and the number of accumulations (Lee and Wang, 1985; Kaufman et al. 1975, Scheunemeyer and Drew, 1983, Baker et al, 1985), or the distribution characteristics of accumulation sizes in a petroleum play (Drew, 1990; Houghton, 1988; Lee, 1993). Little effort has been made to predict undiscovered resource spatial (geographic) distributions. New demands for both better resource management and improved exploration efficiency require a quantitative description of undiscovered petroleum accumulation spatial distribution characteristics (Hood et al, 2000). This provoked a new trend of methodological development aimed at predicting undiscovered petroleum accumulation locations (Meneley et al. 2003). Quantification of the

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 661–670.

undiscovered petroleum resource spatial distribution must simultaneously consider two key elements: the size of and the location of the undiscovered accumulations. Recent studies have attempted to map undiscovered accumulations spatial characteristics using different techniques (e.g., Chen et al, 2000; 2001 and 2002, 2004; Hood *et al.*, 2000, Gao et al. 2000). However, the methods currently available do not allow a full integration of data to account for both, the accumulation size and its location.

Stochastic simulation is a proven tool for revealing uncertainties in petroleum exploitation (e.g., Deutsch, and Tran, 2002; Georgsen *et al.*, 1994; Holden *et al.*, 1998) and could be an ideal tool for undiscovered petroleum accumulation prediction. Its application to petroleum resource assessment could generate equal-probable realizations of potential petroleum accumulation with geographic characteristics. The uncertainties associated with the modeled accumulations provide an important feature for visualizing exploration risk. Currently stochastic simulation considers the spatial correlation characteristics and observational conditions that are derived directly from the exploration drilling results. However, there is no information from drilling results regarding undiscovered petroleum accumulation locations, which presents a major obstacle to the application of stochastic simulation to petroleum resource assessment.

Different geoscience data types contain unique information regarding petroleum accumulation properties. In a mature play at least four data types carry information pertinent to petroleum occurrence spatial characteristics: 1) geological data; 2) exploration drilling results; 3) geophysical data; and 4) location and data quality information regarding geoscience surveys (Chen et al. 2000). Geological information is genetic in character. Available geological information indicates the necessary conditions for petroleum occurrence and it allows, in principle, the inference of petroleum occurrence spatial characteristics (Hood et al. 2000). The spatial variation of geological conditions necessary for petroleum accumulation characterizes the relative favorability for a petroleum accumulation. It is possible to integrate such information and infer the possible locations of undiscovered petroleum accumulations (Chen et al, 2000, 2002).

We propose a model-enhanced stochastic simulation, using a Fourier transform algorithm to solve the problem of missing information related to undiscovered petroleum locations. In the simulation a geological model, in the form of geological favorability or probability of petroleum occurrence derived from the analysis of available geological and geophysical observations, is employed to infer the undiscovered petroleum accumulation locations. We illustrate this approach using an example of the Middle Devonian Rainbow petroleum play in the Western Canada Sedimentary Basin (WCSB). In the example, the pre-1994 exploration data set is used as input. The simulation results are compared against and validated by subsequent exploration (post-1993) drilling results, both successful and unsuccessful.

2 Method Descriptions

2.1 SIMULATION ALGORITHMS

Among different stochastic simulation algorithms, Fourier transform approaches, such as the spectrum simulation algorithm (Pardo-Iguzquiza and Chica-Olmo, 1993) and the phase identification algorithm (Yao, 1999) appear to be ideal for simulating undiscovered petroleum accumulations with geographic references. There are two major advantages in using the Fourier transform method. In addition to the computational advantage, it is possible to study the spatial correlation-specific and location-specific characteristics of the petroleum accumulations separately. Models for spatial correlation and location characteristics can be conveniently extracted from different geoscience data sources and integrated into the simulation.

In a frequency domain, power spectrum, *S(ω)*, is related to covariance function in geostatistics by the Wiener-Khintchine theorem (1993) which states that any stationary process has a covariance function C(h) of the form:

$$C(h) = \int_{-\pi}^{\pi} S(\omega) e^{i\omega h} d\omega \qquad (1)$$

where ω is the angular frequency (*S(ω)* is equal to square of the absolute value of amplitude, $|A(\omega)|^2$). This indicates that both functions (the power spectrum and the covariance) contain the same spatial correlation information, but are expressed in different forms. The spatial correlation is governed by the power spectrum and geographical references are specified by the phase spectrum *φ(ω)*. Inferred locations of undiscovered petroleum accumulation through geological analysis and information integration with geophysical data can be presented in a geological model, such as in the form of a conditional probability map of petroleum occurrence. Such a probability map serves as a spatial density function in the simulation that controls the geographical locations of inferred accumulations. For principles and mathematical formulations of the Fourier transform approach, the reader is referred to Yao (1999).

2.2 A FRACTAL MODEL OF PETROLEUM ACCUMULATION

Barton and Scholz (1995) and La Pointe (1995) studied the data from well-explored petroleum basins and concluded that petroleum accumulation spatial distributions are fractal. La Pointe (1995) proposed using fractal geometry to estimate the total potential in a region. Barton and Scholz (1995) proposed using the fractal dimension as an indicator for exploration planning. Our studies in the Western Canada Sedimentary Basin indicate that petroleum resource spatial distribution exhibits a self-affine characteristic. This characteristic motivated us to examine a fractal model for the quantitative description of petroleum resource spatial distribution.

In the proposed approach, the petroleum resource is described by an image map, on which the value of each pixel represents the average petroleum accumulation magnitude, or yield, within that pixel. The value at each pixel represents the net effect of petroleum accumulation, since the location where petroleum is generated is not necessary the location where it is trapped. Negative values signify net migration away from the pixel, whereas positive values represent net accumulation. The primary objective of petroleum exploration is to find economically recoverable petroleum accumulations and only those accumulations exceeding an economically defined size are significant. We recognize that the economic size threshold can be a variable in time and space, which is primarily dependant on economic and infrastructure conditions. As a result petroleum accumulation spatial distribution patterns may vary with the changes in economics, while the spatial correlation structure remains unaffected.

For a self-similar fractal time series, the power spectrum density has a power law dependency on frequency (Turcotte, 1997, p. 148):

$$S(\varpi) \propto f^{-\beta} \qquad (2)$$

where *f* is frequency*,* and β is an exponential coefficient. In the fractal model, the spatial correlation of objects is fully specified by Eq. (2).

2.3 SIMULATION PROCEDURE

We proposed a simulation that has the following steps:

1) Prepare a petroleum accumulation image map from exploration results;
2) Estimate fractal parameters from the image map according to sampling characteristics;
3) FFT the image map to obtain both amplitude and phase maps;
4) Calibrate the amplitude map using the estimated fractal parameters to obtain a Modified Fractal Amplitude Map (MFAM);
5) Generate a phase map that incorporates information from either geological favorablity or petroleum occurrence probability map;
6) Generate a fractal image (accumulation map) using the MFAM and the inferred phase in step 6 using inverse FFT;
7) Check the fractal image against both exploration observations and geological constraints. Calculate the difference between the simulated values and those at conditioning pixels.
8) If the difference is below the pre-set tolerance, then accept the results;
9) If the difference is greater than the tolerate threshold, modify the accumulation map by replacing the simulated values with the observed values at the conditioning pixels.
10) FFT the modified accumulation map, and get new amplitude and phase maps;
11) Replace the new amplitude map with the MFAM and keep the new phase map in step 10, and inverse FFT the MFAM and new phase. Repeating steps 7) to 10) until a desired tolerance is reached.

3 Application Example

The Rainbow petroleum play, located in northwestern Alberta, WCSB, is a mature exploration play with an areal extent of about 5000 km^2. Major geological controls on this play and its petroleum system are well described (Barss, *et. al*., 1970, Podruski, *et* *al*, 1988, Reinson, *et al*., 1993, Li, *et al*., 1999, Folwer, et al. 2002). Exploration for oil and gas began in this play in the early 1950's. By the end of 1993, 409 wild cats had been drilled, leading to the discoveries of 22 gas pools, 87 oil pools, and 77 oil and gas pools, with a total oil and gas reserve of 269.1x10^6 m^3 (in place) oil equivalent (o.e.). In the subsequent period, from 1994 to 2000, 52 additional exploratory wells were drilled, among which 32 discovered either oil/gas pools or oil/gas flows on test. The pre-94 data set was used to estimate model parameters and to condition the simulation. The post-93 data set served as a test data set to check the predictive value of the simulation output. A discovered petroleum pool map (Fig. 1) was prepared using the Alberta Energy and Utilities Board (EUB) annual reserve report (EUB, 2001). The pool locations are represented by the discovery well locations at the center. Pool size is indicated by the pixel value. A rectangular area of 0.36 km^2 is assumed to be "exhausted" of petroleum potential by an exploratory well. In the simulation, dry wells were used as constraints that excluded petroleum accumulation at the same location.

An amplitude map and a phase map are derived by a Fourier transform of the discovered petroleum accumulations map. Fig. 2 shows the amplitude profiles in easting and northing directions. The deviations of the amplitude from the straight lines are interpreted as exploration bias. The MFAM, calibrated using the fractal parameters β in eq (2), represents the spatial correlation for all petroleum accumulations in the size range indicated by the data. The phase map derived from the discovered petroleum accumulations contains no information regarding the locations of the undiscovered petroleum accumulations. The use of the traditional conditional simulation with the amplitude map results in a random realization of a stochastic simulation conditioned on both the discovered petroleum accumulations and the dry well locations. In such a realization, the spatial correlation in the amplitude map is retained, but the locations of the undiscovered petroleum pools could be anywhere except at the conditioned pixels.

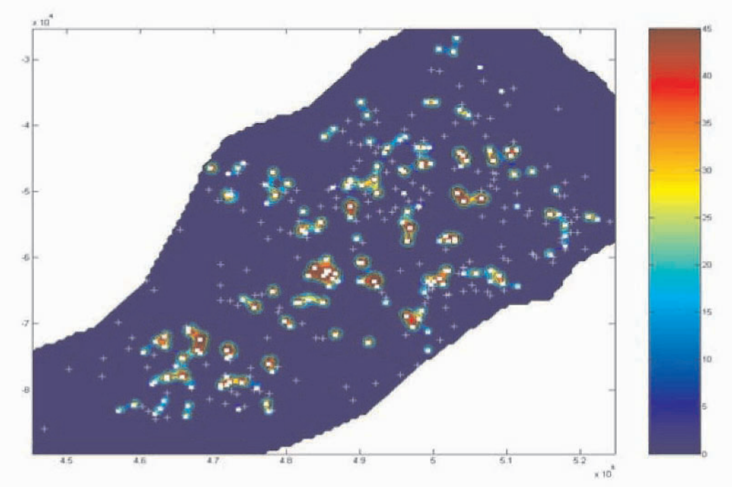


Figure 1 Pre-1994 discoveries (crude oil and natural gas) in the Rainbow play, WCSB. Open squares indicate the discovery well locations, and crosses indicate dry well locations.

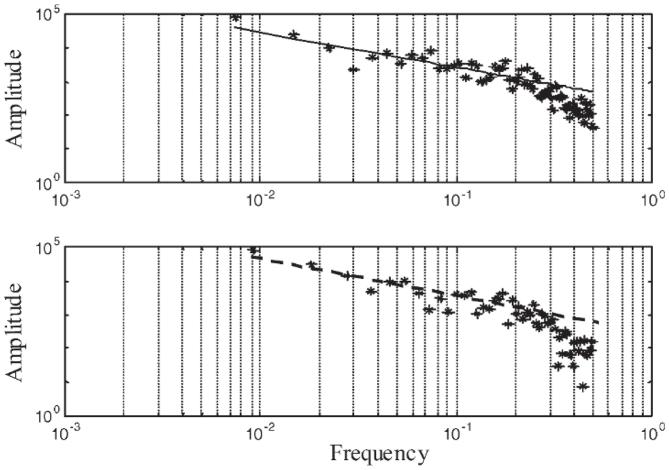


Figure 2 Power spectrum profiles in x (above) and y (below) directions show the under-sampling of the smaller petroleum accumulations indicated by the deviation of the linear relationships in higher frequency regions (horizontal axis: frequency and vertical axis: amplitude).

In the same area a previous study of petroleum accumulation spatial distribution characteristics resulted in a conditional probability map of petroleum occurrence (Chen et al. 2001). That map integrates information from geological factors describing petroleum occurrences. With this independently determined conditional probability map, the iteratively repeated simulation procedure produces equal-probable realizations, which are validated against not only, discovered petroleum accumulations, dry wells, and the exhaustion of potential by previous activity, but also the geological conditions controlling petroleum accumulations. Fig. 3 is a probability map of petroleum occurrence based on 3000 conditional realizations. It represents the uncertainty associated with the predicted undiscovered accumulation locations in the play. The likely sizes of both discovered and undiscovered petroleum accumulations, with geographic significance, as predicted by the model-enhanced simulation, are presented in Fig. 4.

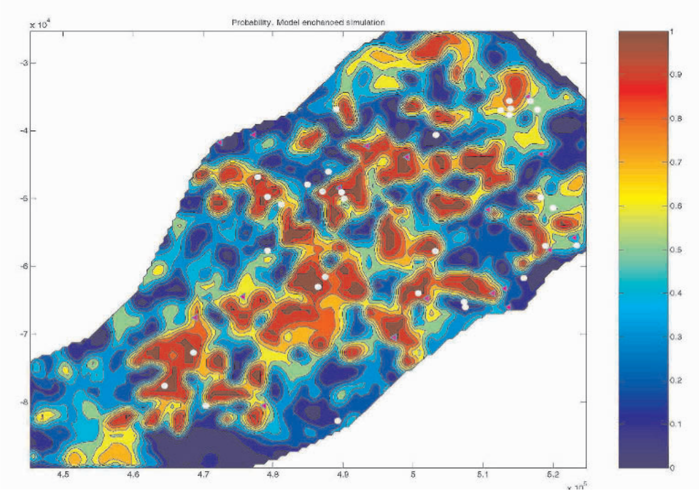


Figure 3 A probability map of petroleum occurrence for the Rainbow play based on 3000 realizations from the model enhanced simulation. Solid circles locate successful exploratory wells and triangles locate unsuccessful exploratory wells locations drilled between 1994 and 2000.

4 Discussion and conclusions

The model-enhanced approach produces a petroleum accumulation map, on which both the size and location of the undiscovered petroleum accumulations are predicted. A probability map from multiple realizations of the simulation highlights the areas with low and high probability values, providing a general view of the exploration risk. Comparison of the post-1993 discovery wells locations with the resulting predicted petroleum occurrence probability validates this approach. Twenty-two of the thirty-two post-1993 discovery wells are located in areas with predicted probability values >0.5. Fifteen of these well locations occur where probability values are >0.7. The simulations have resulted in relatively high petroleum occurrence probabilities in a less explored area in the northeast part of the play, where only one unsuccessful well was drilled prior to 1994. Seven post-1993 exploratory wells were completed in this part of the play. Six were discoveries. The proposed method produces maps (probability and resource maps, Figs. 3 and 4) showing a profound influence of the geological characteristics of the play. In contrast, the traditional conditional simulation, without using the additional geological information, did not predict the six post-1993 discoveries in the northeastern part of the play and produces a more random pattern in areas without well control. This suggests that the proposed approach captures the essentials of the petroleum accumulation spatial features, as well as, the their geological characteristics.

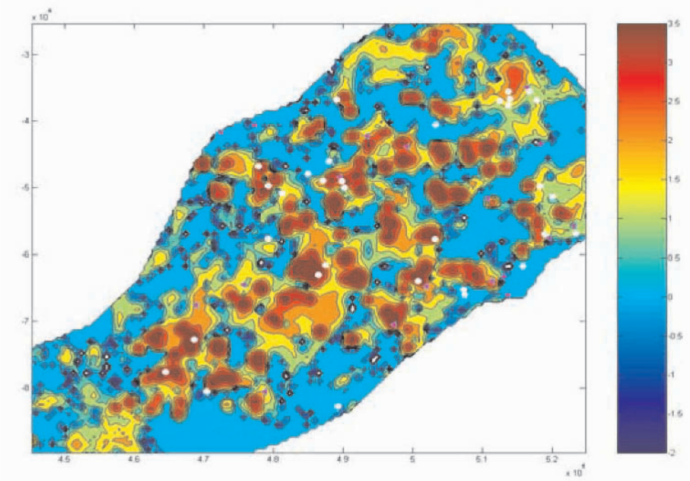


Figure 4 Predicted sizes (logarithmic value) of petroleum accumulations from the model-enhanced simulation. The average value of 3000 realizations represents the size at each pixel. Solid circles locate successful exploratory wells and triangles locate unsuccessful exploratory wells locations drilled between 1994 and 2000.

We have demonstrated using the Rainbow petroleum play example that the use of additional geological and geophysical prospecting data enhances the spatial modeling by contributing location-specific information to the phase map. If resource location is an important feature in the analysis, Fourier transform algorithm is an ideal approach for the conditional simulation, because it allows the spatial correlation structure and location-specific information from different sources to be analyzed separately in a frequency domain and integrated in a spatial domain.

Acknowledgements

This work was supported by Geological Survey of Canada (GSC) and the Panel for Energy Research and Development, NRCAN. The authors thank Drs. Glen Stockmal, Dale Issler and Maowen Li at GSC Calgary for their helpful discussions, and Ms. Ping Tzeng and Dr. David Lepard at GSC Calgary for the help with data recovery and coordinates conversions. GSC contribution # 2004066.

References

Baker, R.A, Gehman, H.M., James, W.R., and White, D.A., Geological field number and size assessment of oil and gas plays: in *Oil and Gas Assessment - Methods and Applications*, edited by D. D. Rice; AAPG Studies in Geology #21, 1986, p.25-32.

Barss, D.L., Copland, A. B., and Retch, W. D., Geology of Middle Devonian reefs, Rainbow area, Alberta, Canada; in M. T. Halbouty ed. Geology of giant petroleum fields, AAPG Memoir 14, 1970, p. 9-49.

Barton, C. C. and Scholz, C. H., The fractal size and spatial distribution of petroleum accumulations, in Fractal in Petroleum Geology and Earth Processes, edited by Barton and La Pointe, 1995, Plenum Press, New York, 1995, p13-34.

Chen, Z., Osadetz, K., Gao, H., Hannigan, P. and Watson, C., Characterizing the spatial distribution of an undiscovered petroleum resource: the Keg River Reef play, Western Canada Sedimentary Basin, Bulletin of Canadian Petroleum Geology, v.48, no.2, 2000, p.150-163.

Chen, Z., Osadetz, K., Gao, H., and Hannigan, P., Improving exploration success through uncertainty mapping, the Keg River reef play, Western Canada Sedimentary Basin, Bulletin of Canadian Petroleum Geology, v. 49, no. 3, 2001, p.367-375.

Chen, Z., Osadetz, K. A. Embry, and Hannigan, P., Geological favorability mapping of petroleum potential using fuzzy integration, example from western Sverdrup Basin, Canadian Arctic Archipelago, Bulletin of Canadian Petroleum Geology, v50. no. 4, 2002, p.492-506.

Chen, Z., Osadetz, K.G., Gao, H. and Hannigan,P., SuperSD: an object-based stochastic simulation program for modeling locations of undiscovered petroleum accumulations, Computer and Geosciences, v. 30, 2004, p.281-290.

Deutsch, C. V. and Tran, T. T., FLUVSIM: a program for object-based stochastic modeling of fluvial depositional systems, Computers & Geosciences 28 (4), 2002, p.525-535.

Drew, L.J., 1990, Oil and gas forecasting, reflections of a petroleum geologist, New York, Oxford University press, International Association for Mathematical Geology, Studies in Mathematical Geology #2.

EUB, 2001, Alberta's reserves 2000 and supply/demand outlook 2001-2010, Statistical series 2001-98, Alberta Energy and Utilities Board, Calgary Alberta.

Fowler, M.G., Stasiuk, L. D., Hearn, M. and Obwemajer, M., Devonian petroleum source rocks and their derived oils in the Western Canada Sedimentary Basin, Bulletin of Canadian Petroleum Geology, v.49, no. 1, 2002, p. 117-148.

Gao, Haiyu, Chen, Zhuoheng, Osadetz, Kirk, and Hannigan, Peter, A pool-based model of the spatial distribution of undiscovered petroleum resources, Math. Geology, v.32, n.6, 2000, p.725-749

Georgsen, F., Egeland, T., Knarud, R., and Omre, H., Conditional simulation of facies architecture in fluvial reservoirs, In: Armstrong M. and Dowd, P. A. (eds.), Geostatistical Simulation, v. 7 of Quantitative Geology and Geostatistics, Proceedings of the Geostatistical Simulation Workshop, Fontainebleau, France, 1993: Kluwer Academic Publ., Dordrecht, 1994, p.235-250.

Holden, L., Hauge, R., Skare, O., and Skorstad, A., Modeling of fluvial reservoirs with object models: Mathematical Geology, 30(5), 1998, p.473-496.

Hood, B.C., South,B.C.,Walton,F.D. and Baldwin,O.D., Use of geographic information systems in petroleum resource assessment and opportunity analysis, in T.C. Coburm and J.M. Yarus, eds., Geographic information systems in petroleum exploration and development: AAPG Computer Applications in Geology, No. 4, 2000, p.173-186.

Houghton, J. C., Use of the truncated shifted Pareto distribution in assessing size distributions of oil and gas fields: Mathematical Geology, v. 20, no. 8, 1988, p. 907-938.

Kaufman, G. M., Balcer, Y. and Kruyt, D., A probabilistic model of oil and gas discovery, in J. Haun, ed., Estimating the volume of undiscovered oil and gas resources: AAPG Studies in Geology Series no. 1, 1975, p.113-142.

La Pointe, P. R., Estimation of undiscovered petroleum potential through fractal geometry, in Fractal in Geology and Earth Sciences, edited by Christopher C. Barton and paul R. La Pointe, Plenume Press, New York, 1995, p35-57.

Lee, P. J., Oil and gas size probability distributions, J-shaped, lognormal, or Pareto?: GSC Current Research, 1993, p. 93-1.

Lee, P. J., and Wang, P. C. C., Prediction of oil or gas pool sizes when discovery record is available: Math. Geology, v. 17, no. 2, 1985, p. 95-113.

Li, Maowen, Fowler, M.G., Obwemajer, M., Stasiuk, L.D. and Snowdon, L. R., Geochemical chacterisation of middle Devonian oils in NW Alberta, Canada, possible source and maturity effect on pyrrolic nitrogen compounds, Organic geochemistry, v. 30, 1999, p1039-1057.

Meneley, R., Calverley, A. E., Logan,K.G. and Procter,R.M., Resource assessment methodologies: Current Status and future directions, AAPG Bulletin, v87, n.4, 2003, p.535-540.

Pardo-Iguzquiza, E. and Chica-Olmo, M., The Fourier integral method: an efficient spectral method for simulation of random fields: Math. Geology, v.25, no. 4, 1993, p.177-217.

Podruski, J. A., Barclay, J. E., Hamblin, A. P., Lee, P J., Osedetz, K. G., Proctor, R. M. and Taylor, G. C., Conventional oil resources of western Canada, Geological Survey of Canada Paper 87-26. 1988, P.20-27

Reinson, G.E., Lee, P.J., Warters, W., Osadetz, K.G., Bell, L.L., Price, P. R., Trollope, F., Campbell, R. I., and Barclay, J.E., 1993, Devonian gas resources of the Western Canada Sedimentary Basin, Geological Survey of Canada Bulletin 452.

Schuenemeyer, J. H. and Drew, L. J., A procedure to estimate the parent population of the size of oil and gas fields as revealed by a study of economic truncation: Mathematical Geology, v. 15, no 1, 1983, p. 145-162.

Turcotte, D.L., Fractals and chaos in geology and geophysics, 2nd edition, Cambridge University Press, 1994, p.148.

Yao, T., Conditional spectral simulation with phase identification: Math. Geology, v. 30, no. 3, 1999, p. 285-308.

PREDICTION OF SPATIAL PATTERNS OF SATURATION TIME-LAPSE FROM TIME-LAPSE SEISMIC

JIANBING WU[1], TAPAN MUKERJI[2] and ANDRE G. JOURNEL[1]
[1]*Department of Petroleum Engineering, Green Earth Science Building*
Stanford University, USA, 94305
[2]*Department of Geophysics, Mitchell Earth Science Building*
Stanford University, USA, 94305

Abstract. The advent of 4D time-lapse seismic surveys opens a new dimension to reservoir development: the possibility of monitoring from seismic data fluid flow hence production. Because of the low resolution of seismic data, one should not expect any useful point-to-point correlation between time-lapse saturation and seismic data; instead one might expect correlation of spatial patterns of these 2 variables. Spatial patterns involve multiple-locations within a fixed template window, and are summarized by the principal/canonical components of the within-template variability of each variable.

1 Introduction

4D seismic surveys, possibly with permanent downhole sensors, are being considered to monitor fluid production through observing changes in reservoir state. Time differences of seismic attributes are related to changes in pore fluids and pore pressure because bulk density and bulk moduli change during the drainage of the reservoir. Maps of seismic time difference can be used to detect fingering, monitor fluid movement, improve recovery and locate new wells (Nur, 1989; Anderson, 1998; Lumley, 1999). Clear success stories are presently limited to clastic reservoirs, shallow reservoirs and reservoirs where gas flow allows a greater density differentiation. There have also been some successful applications to carbonate reservoir (Hirsche, 1997; Talley, 1998). In the best case, correctly processed time-lapse seismic data can point out to fluid movement through mere visual inspection without any need for correlation statistics. These clear success stories may have led to dismissing the potential of 4D seismic surveys in less favorable cases. In such unfavorable cases, there may still be some influence of fluid saturation changes on the seismic data, but detection of such weak relation would need filtering and correlation tools beyond mere visual inspection (Sønneland, 1997).

A wider utilization of 4D seismic surveys for monitoring production would have to settle with lesser expectation: water or gas fronts may not be seen deterministically, yet may have a detectable influence, if only tenuous and stochastic, on the seismic attributes. Time-lapse seismic data could then be systematically considered as a covariate data whose correlation would vary from quasi perfect (present best cases) to none. The intermediary cases are the object of this study.

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 671–680.

A preliminary requisite to using any covariate data is to establish its correlation with the primary variable being estimated. There are, however, many ways one spatially distributed variable, say $y(\mathbf{u})$, may be dependent on another one, say $z(\mathbf{u})$, besides the trivial point-to-point correlation calling for the 2 variables to be co-located at the same location of coordinates vector $\mathbf{u}$:

– first, the 2 variables may be defined on different volume supports and have different space and time resolution. Typically, the vertical resolution of seismic data is lesser than the vertical discretization of numerical reservoir models, those used to predict flow movement.

– it may be that only specific spatial patterns of the covariate $y(\mathbf{u})$ carry a relation with either the primary variable $z(\mathbf{u})$ itself or some of its z-spatial patterns.

The covariate $y(\mathbf{u})$ may be a time difference of seismic amplitudes measured at location $\mathbf{u}$, $z(\mathbf{u})$ could be the corresponding time difference in water saturation defined over a flow simulator block co-centred at $\mathbf{u}$. The potentially valuable time-lapse seismic information $y(\mathbf{u})$ should not be dismissed just because, from a few calibration wells, the co-located correlation { $y(\mathbf{u})$, $z(\mathbf{u})$ } was found to be poor or very poor.

2 Setting the experiment: The Stanford V reservoir

The Stanford V reservoir is a large 3D synthetic data set modeling a clastic reservoir made up of meandering fluvial channels with crevasse splays and levies in a mud background (Mao, 1999). The second layer of Stanford V is retained here as the reference reservoir, with a net to gross ratio of 0.53. This reference reservoir is discretized by a 3D grid with 100×130×10 nodes. Figure 1 shows a schematic vertical cross section of that layer (right figure), and its 3D facies distribution.

2.1 FLOW SIMULATION

To maximize sensitivity of the seismic time-lapse data, the reservoir is assumed shallow (top depth at 600m, see Figure 1), with light oil density at 45API°. The initial water saturation is 0.15 in sandstone and crevasse, and 0.30 in mudstone, corresponding to a water-wet mudstone that, globally, contributes a substantial amount of oil.

One injector is located in the SW corner at grid node (10,10), and one producer in the NE corner at grid node (90, 120). The water injection rate is 40,000 STB/day. During production no gas is emitted from the oil phase. The Eclipse simulator was run for water flooding over a total period of 20 years starting Jan.1, 2000. Breakthrough occurs at the end of 2013. Figure 2 gives the water saturation on layers 5-7 on Dec.29, 2013 just before breakthrough.

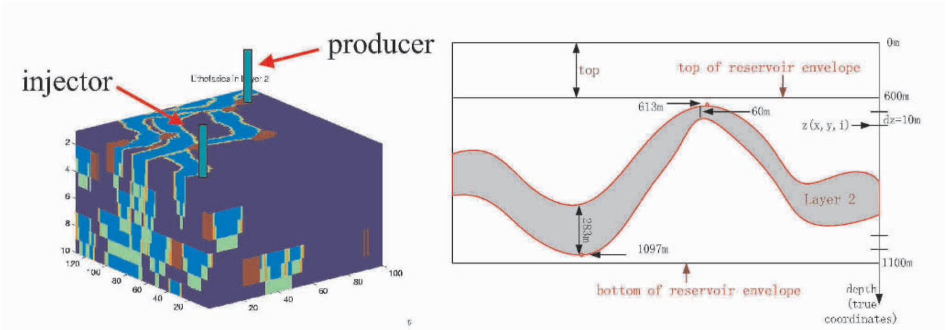


Figure 1. 3D Facies distribution (left) and a schematic vertical cross section of the reference reservoir.

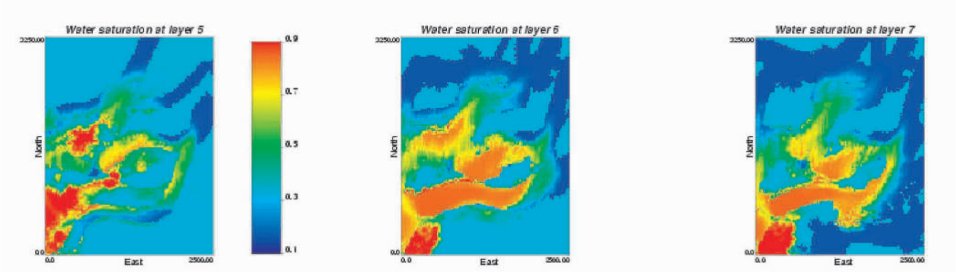


Figure 2. Water saturation on stratigraphic layers 5-7 on Dec. 29, 2013.

2.2 SEISMIC SIMULATION

After obtaining the saturation and pressure from flow simulation, the amplitude seismic traces were forward simulated using a normal incidence 1D convolution model with Fresnel zone lateral averaging, see Wu (2003) for greater details.

This seismic data simulation was repeated at different times during the 20 years production period to mimic 4D surveys attempting to track the changes in water (brine) saturation.

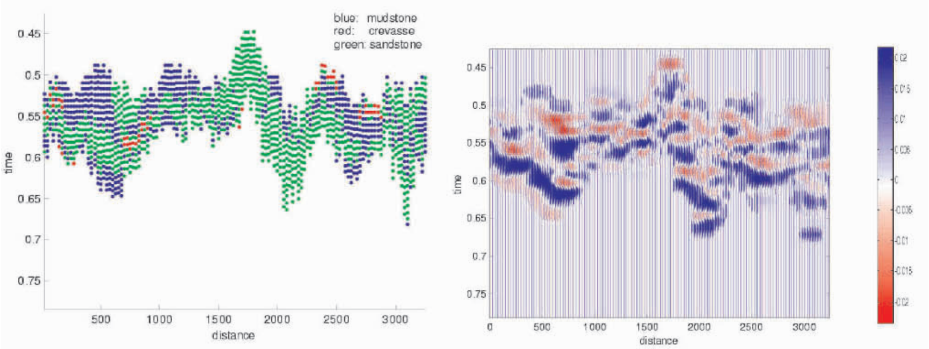


Figure 3. Facies distribution (left) in NS vertical section at x=1, seismic amplitude (right) in the same section at the initial time

Figure 3 gives the 3 facies distribution (channel, crevasse, mud) over the first NS vertical section at x=1 (left figure), and the seismic amplitude in the same section at the initial time Jan.1, 2000. From Figure 3 it would be difficult to retrieve the facies distribution, the reason being that seismic amplitudes vary within the same facies because of the within-facies petrophysical variability. Figure 4 shows the initial seismic amplitude at layers 5-7.

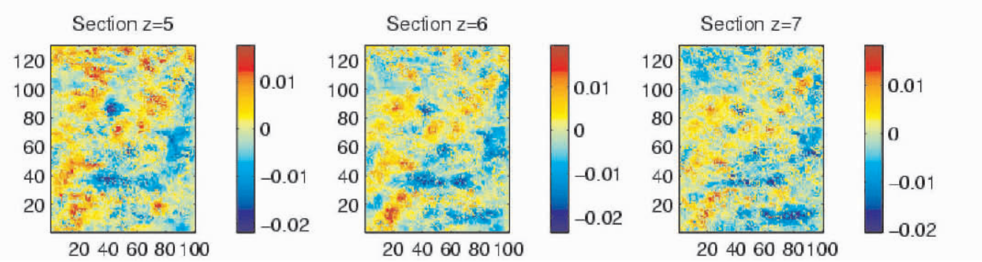


Figure 4. Initial seismic amplitude at stratigraphic layers 5-7

3 Point-to-point correlation

We attempted a direct point-to-point correlation between the two previously generated seismic amplitude and water saturation fields on Jan.1, 2000. Not surprisingly, that correlation came out very low at -0.06.

Indeed seismic amplitude and water saturation are defined on very different volume supports. Seismic amplitude is an average of reflectivity coefficients over a large horizontal Fresnel zone, here of dimension 9×9 grid nodes, i.e. of size 225m×225m in average. In addition, seismic amplitude records vertical impedance contrast. Consequently, we would expect better correlation between seismic amplitude and a vertical contrast of spatially averaged saturation values:

- average first the water saturation data over a 3D moving window approximating the seismic Fresnel zone. That average is porosity-weighted:

$$\overline{Sw}(\mathbf{u}) = \frac{\sum_{l-1}^{L} \phi_l \times Sw_l}{\sum_{l-1}^{L} \phi_l} \qquad (1)$$

where L=243 is the number of grid nodes of the 9×9×3 window centred at location $\mathbf{u} = (i, j, k)$ in stratigraphic coordinates; ϕ_l and Sw_l are the node l porosity and saturation values.

- next, define the vertical saturation contrast as:

$$d\overline{Sw}(i, j,k) = \overline{Sw}(i, j,k) - \overline{Sw}(i, j,k-1) \qquad (2)$$

where k is the vertical stratigraphic coordinate, increasing with depth.

Figure 5 gives the water saturation vertical contrast on layers 5-7. The overall 3D collocated correlation between that Figure 5 and Figure 4 is now -0.20, slightly better than the previous -0.06.

However that -0.2 correlation still does not reflect the visual patterns correlation seen on the Figures 4 and 5.

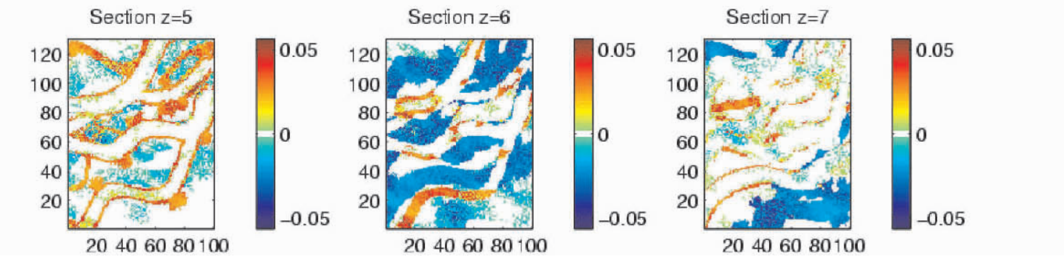


Figure 5. Initial water saturation vertical contrast at stratigraphic layers 5-7

4 Spatial pattern correlation

Because the point-to-point correlation (-0.20) does not render justice to the visual (pattern) correlation seen between Figures 4 and 5, we have to define a better correlation tool, yet one which is not too case-specific and could be applied to a whole range of reservoir heterogeneities and seismic surveys.

The goal here is not so much to detect spatial patterns within a 3D seismic cube, but to define a measure able to correlate fuzzy spatial patterns between 2 cubes, the seismic cube and the water saturation one. There are many classification tools (Duda, 2001), but their primary goal is to detect patterns independently of correlation. As an initial choice we have focused on principal component analysis (PCA) and canonical analysis (CA) (Michael, 1984; Jolliffe, 1986).

The general idea of CA is to define spatial templates, one for seismic data, one for saturation data, then to define within each template a linear combination of the data which would maximizes the cross-correlation seismic vs. saturation. PCA has the added advantage that each linear combination contributes maximally to the within-template variance.

4.1 TEMPLATE SIZE

Because seismic data are already averaged over a horizontal Fresnel zone, here 9×9 grid nodes corresponding to 225m×225m in average, the seismic template needs only extend vertically. In depth coordinate we retain a vertical column template of size (1×1)×7 centred on each grid node location u (Figure 6: left)

To approximate the seismic Fresnel zone, we consider for saturation data a full 3D template of size (5×5)×3, see right of Figure6. To reduce the actual dimension of these templates, the 25 saturation data of each of its horizon sections are averaged into 3 values: the central value, the average of the 8 first aureola values (labelled by '×'), the average of 16 outer aureola values (labelled by 'o'). Thus a saturation template comprises 3×3 = 9 values, as opposed to 7 values for the seismic template.

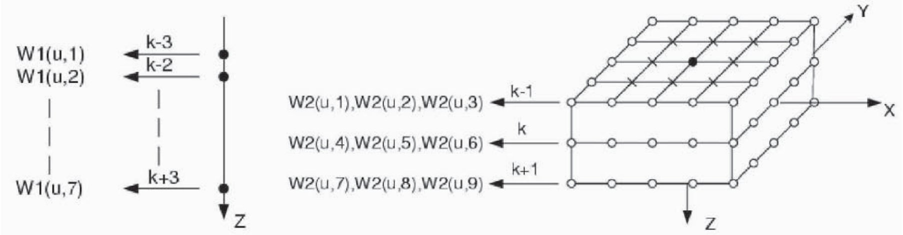


Figure 6. Seismic template (left) and water saturation template (right)

4.2 DATA TABLES

The two 3D cubes of seismic and saturation data are scanned with the two previous templates generating two large data tables,
– of size N rows × 7 columns for seismic data
– of size N × 9 for saturation variable, where N is the number of template centres. Here N=477,310.

The mean and variance of each column of these tables are calculated and the corresponding column values are standardized to mean zero and unit variance. The covariance matrix is calculated from each standardized data table; that matrix is of size (7×7) for seismic, (9×9) for saturation. Finally, PCA and CA is performed using these covariance matrices.

4.3 APPLICATION TO 3D CORRELATION

PCA was applied to the data recorded on Jan.1, 2000, more precisely the 3D cube of seismic amplitude $seis(i,j,k)$ and the corresponding 3D cube of vertical saturation difference $dSw(i,j,k) = Sw(i,j,k) - Sw(i,j,k-1)$, see Figures 7 and 8.

The 1st seismic PC explains 84% of the within-template variance, and the 1st saturation difference PC explains only 59% of its template variance. The overall 3D point-to-point correlation between these two first PC is 0.39, a value still low but more reflective of the patterns correlation seen between Figures 4 and 5.

The same PCA was repeated on the data recorded on Jan.1, 2004. The resulting 3D point-to-point correlation between the first PC's of seismic amplitude and water saturation vertical difference was found to be 0.32, an equally not too high value.

In preliminary conclusion, it appears that PCA succeeds, to a limited degree, to recognize some of the patterns correlation existing between seismic attribute and vertical difference of spatially averaged water saturation. That observation should be now checked on time difference of both seismic and saturation data.

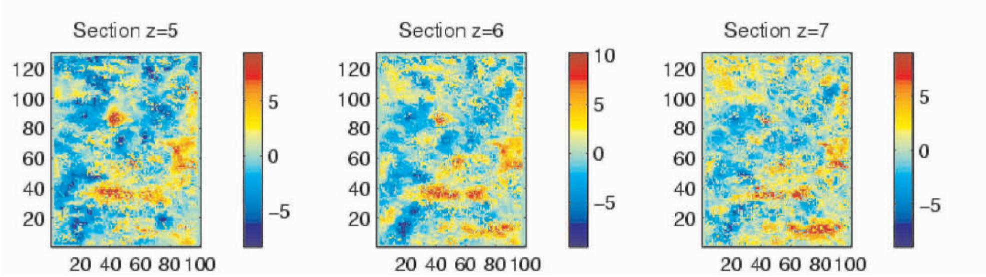


***Figure** 7*. First PC values of seismic amplitude at stratigraphic layers 5-7 (Jan. 1 2000)

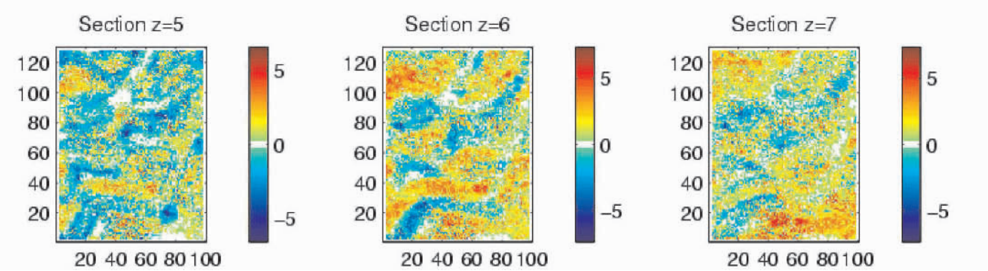


Figure 8. First PC values of water saturation vertical contrast at stratigraphic layers 5-7 (Jan. 1 2000)

4.4 APPLICATION TO 4D CORRELATION

The previous PCA and CA were repeated on now time-lapse data. More precisely:

- the 3D seismic data used is now a time-lapse of seismic amplitude:

$$\delta_{seis}(i,j,k,t_2,t_1) = seis(i,j,k,t_2) - seis(i,j,k,t_1) \qquad (3)$$

- the 3D saturation data is the corresponding time difference of water saturation vertical difference:

$$\Delta_{sw}(i,j,k,t_2,t_1) = dSw(i,j,k,t_2) - dSw(i,j,k,t_1) \qquad (4)$$

with $dSw(i,j,k) = Sw(i,j,k) - Sw(i,j,k-1)$.

The time lapse here considered is: t2 =Jan.1, 2004, t1 =Jan.1, 2000.

The 1st seismic PC explained 53% of the within-template variance, and the 1st saturation PC explains 48% of its template variance. The overall 4D (actually 3D for time difference) correlation between these two first PC's is 0.78. This large correlation value is a good omen for using time lapse seismic to monitor water saturation changes. Figures 9 and 10 gives the maps of these two first PC's values over the 9 stratigraphic reservoir surfaces, these maps clearly show high visual correlations between the seismic time lapse and water saturation difference time lapse.

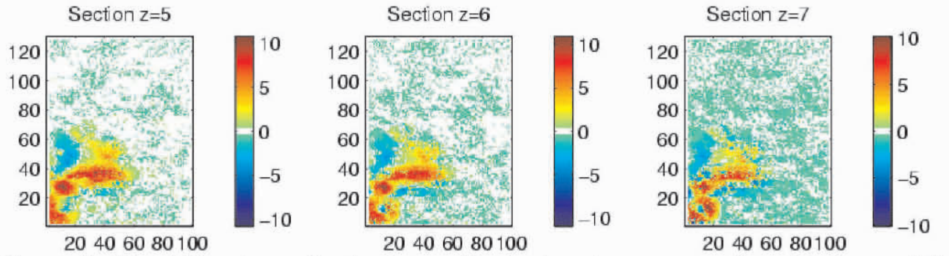


Figure 9. First PC values of seismic amplitude time lapse at stratigraphic layers 5-7 (Jan. 1 2000 – Jan. 1 2004)

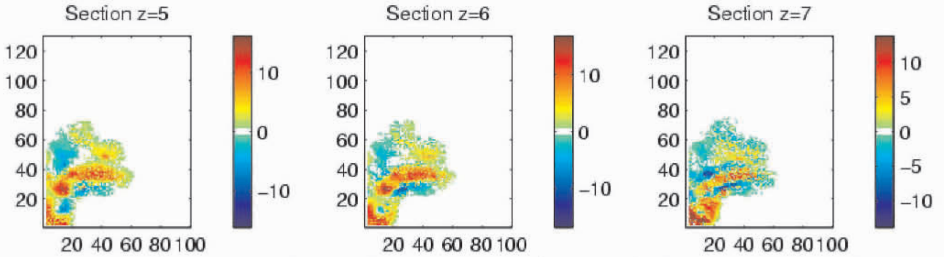


Figure 10. First PC values of water saturation vertical contrast time lapse at stratigraphic layers 5-7 (Jan. 1 2000 – Jan. 1 2004)

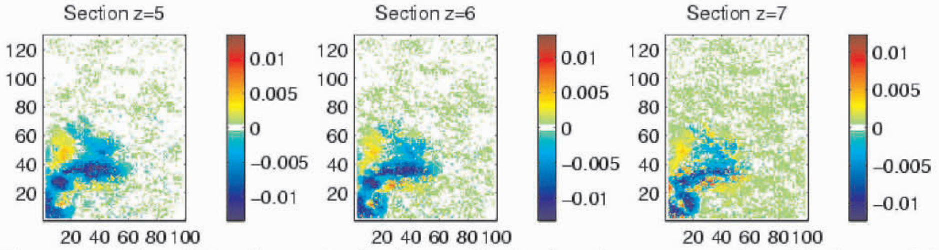


Figure 11. First CC values of seismic amplitude time lapse at stratigraphic layers 5-7 (Jan. 1 2000 – Jan. 1 2004)

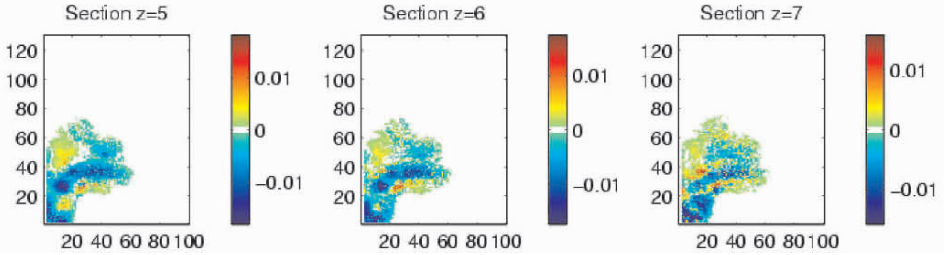


Figure 12. First CC values of water saturation vertical contrast time lapse at stratigraphic layers 5-7 (Jan. 1 2000 – Jan. 1 2004)

Next, canonical analysis (CA) was applied to the two previous sets of time-lapse data. The resulting first pair of canonical components, $CC_{seis}(\mathbf{u})$ and $CC_{sw}(\mathbf{u})$, features a high cross correlation of 0.82. The corresponding two series of CC maps are shown on Figures 11 and 12.

5 Conclusions

Because of their different resolution (different support volumes) the point-to-point correlation between saturations and seismic time lapse might be low, although there may be significant correspondence between spatial patterns of the two time-lapse variables. A spatial pattern, although immediately visible to the eye, is a complex statistical concept involving multiple-points in space. Correlation between multiple-point data events is not yet well understood nor does it exist any established measure for it.

The idea is to summarize a multiple-point data event, as defined within a fixed template of points, by a few linear combinations of these point values. Traditional correlation measures can then be applied to such linear combinations. The linear combinations should be indicative of the within-template spatial patterns, and should display significant cross-correlation between saturations and seismic time lapse variables. The linear combinations provided by principal component analysis (PCA) comes to mind: by definition, the first few principal components (PC's) or linear combinations explain a large part of the within-template spatial variance. Because of that large variance contribution, it is conjectured that PC's are good summaries of potential spatial patterns existing within the template area. Another, more direct, approach is canonical analysis (CA), which seeks at determining the two linear combinations (one for saturation, one for seismic time lapse) with maximum cross-correlation independently of their respective within-template variance contributions.

Both PCA and CA were applied to the synthetic clastic Stanford V reservoir, where both seismic and reference water saturation time-lapse data are available and pattern correspondences are visually evident. The two sets of data (seismic and saturation) were analyzed through the filters of their respective templates, vertical for seismic, 3D mimicking the Fresnel zone for saturation. Although the original point-to-point correlation between the two time-lapse variables was insignificant around 0.1, correlation of the two first PC's (one for seismic one for saturation) is high around 0.7, a value more reflective of the excellent visual patterns correspondence. Canonical analysis increases that correlation up to almost 0.8.

In real practice, such proxies for spatial patterns correlation could be established from simulated reservoirs on which 4D seismic is forward simulated. Such correlation could be used to estimate or simulate time-lapse saturation values using observed PC's or CC's values of seismic time-lapse data. Geostatistics provides tools for estimating or simulating a variable such as water saturation, conditional to linear combinations of another variable (seismic time lapse data). This aspect of the study can now be addressed, because we have shown that, indeed, principal and/or canonical components do carry multiple-point pattern information.

Acknowledgements

We would like to acknowledge the industry affiliates of Stanford Center for Reservoir Forecasting (SCRF).

References

Anderson, R., Guerin G., He, W., Boulanger, A., Mello, U., & Watson, T., 4-D seismic reservoir simulation in a South Timbalier 295 turbidite reservoir, The Leading Edge, October 1998, pp. 1416-1418.

Duda, R., Hart, P., & Stork, D., Pattern Classification, John Wiley & Sons, Inc., 2001

Hirsche, K., Batzle, M., Knight, R.,Wang, Z.J.,Mewhort, L., Davis, R. & Sedgwick G., Seismic monitoring of gas floods in carbonate reservoirs: From rock physics to field testing, SEG expanded abstracts, 1997, pp. 902-905.

Jackson S., Cluster Analysis and Forward Seismic Modelling ofMarine Seismic Data, Stanford University, 1986

Jolliffe I.T., Principal Component Analysis, Springer-Verlag, New York, 1986

Lumley, D.E., Numms, A.G., Delorme, G. & Bee, M.F., Meren Field, Nigeria: A 4D Seismic Case Study, SEG Expanded Abstracts, 1999, pp. 1628-1631

Mao, S., Multiple Layer Surface Mapping with Seismic Data and Well Data, PhD thesis, Stanford University, Stanford, CA., 1999

Nur, A., Direct detection of hydrocarbons and 4D seismology: the petrophysical basis, Burkhardt, H., chairperson. (Technical Programme and Abstracts of Papers – European Association of Exploration Geophysicists: 1989, Vol. 51, pp. 2)

Michael J., Theory and applications of correspondence analysis, Orlando, Fla. Academic Press, 1984

Sønneland, L., Veire, H.H., Raymond, B., Signer, C., Pedersen, L., Ryan, S., & Sayers, C., Seismic reservoir monitoring on Gullfaks, The Leading Edge, 1997 September, pp. 1247-1252.

Talley, D.J., Davis, T.L., Benson, R.D. & Roche, S.L., Dynamic reservoir characterization of Vacuum Field, The Leading Edge, 1998, pp. 1396-1402.

Wu, J.B., 4D seismic amplitude applied to water control on Stanford V, MS thesis, Stanford University, Stanford, CA, 2003

STATISTICAL SCALE-UP: CONCEPTS AND APPLICATION TO RESERVOIR FLOW SIMULATION PRACTICE

LARRY W. LAKE, SANJAY SRINIVASAN AND ABRAHAM JOHN
Department of Petroleum & Geosystems Engineering
University of Texas at Austin, United States, 78712-0228

Abstract

Petrophysical measurements are used to construct reservoir models at a scale that are different from that at which they are measured. This disparity necessitates an adjustment or scale-up of the measured values before they are used. Scale-up is complicated by the properties being heterogeneously distributed in space and self- or autocorrelated. The autocorrelation means that the heterogeneity itself must be preserved during the scale up.

We discuss scale up in the following two contexts. The first is the scaling of permeability and ultimate recovery efficiency as the size of a flow field increases. The second is the nature of adjustments to the model properties that are needed to better reconcile the observed behavior at different scales. We look at the effect of scale up procedures on the distribution and correlation structure of fluid velocities. Despite the long history of scaling up reservoir simulation models, we find relatively little literature on this type of scaling up.

Motivation

Figure 1 shows how lateral (horizontal or bed parallel) permeability increases with scale. Among the most common changes made during a history matching procedure is that core or log-derived permeability must usually be increased to match field performance data. This increase is expected because there are formations from which fluids are readily produced, but cores from which, being dominated by small scale heterogeneity, will pass little fluid. The same scale effect is also seen in vertical (bed normal) permeability except that it *decreases* with scale (Lake and Srinivasan, 2004).

Another scale effect is in the dependence of ultimate hydrocarbon recovery efficiency on size in enhanced oil recovery or remediation projects. This is shown in Fig. 2. Despite the scatter of points (largely caused by differences in process type), the ultimate recovery efficiency decreases with increasing volume or scale. The decrease shown in Fig. 2 is undoubtedly one of the reasons for the slow acceptance of advanced recovery processes. A similar (but opposite in trend) phenomenon occurs in the scale dependence of dispersivitiy, which has been shown in several works (Mahadavan et al., 2003, for example).

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 681–690.

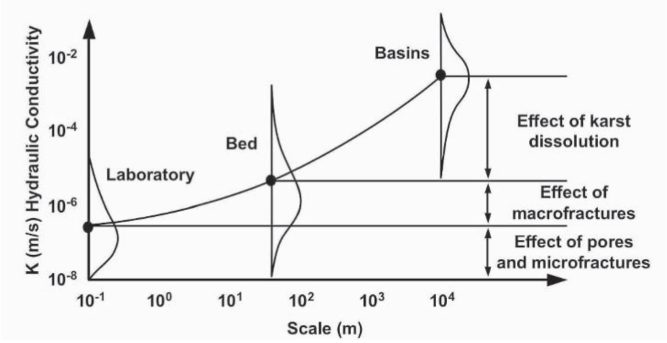


Figure 1. Effect of scale on lateral permeability and its distribution. Adapted from Kiraly (1975).

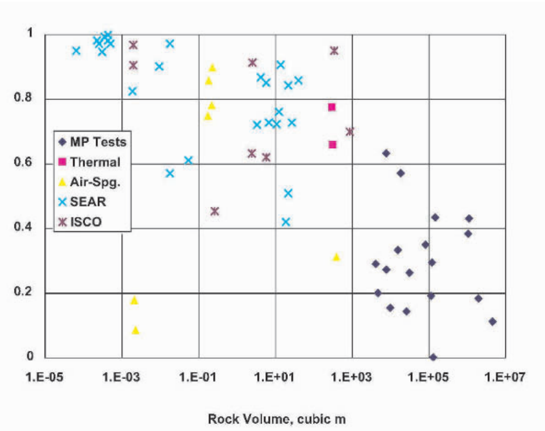


Figure 2: Plot of ultimate hydrocarbon recoveries versus scale. All points from Soga et al., 2004, except micellar-polymer (MP) points, which are from Lake and Pope, 1978.

The observations in Figs.1 and 2 may be explained using geology, the vertical permeability effect, in particular, by the presence of partially permeable or impermeable discontinuous shales. While this is undoubtedly so, such behavior can also be explained statistically. It is demonstrated in this paper that the increasing trend in horizontal permeability as well as the decreasing trend in k_V/k_H ratio with scale can be explained in terms of the dispersion variance or the variance of the mean. In other words, it is argued that the observed trends in permeability can be explained in terms of scaling of reservoir heterogeneity.

The scale behavior of ultimate recovery and dispersivity depend on the heterogeneity of the resulting velocity field. We discuss this in the last part of this paper. For the most part, a more heterogeneous velocity distribution will result in a smaller ultimate recovery and a larger dispersivity.

Average Permeability

The behavior of average lateral and vertical permeability is explained using one-dimensional analytical statistics. Let *Z* be a scalar, spatially continuous, Gaussian, random variable distributed in one-dimensional field. The average of Z over a distance L has a variance:

$$Var\left(\overline{Z}\right) = \frac{2\sigma^2}{L^2}\left(\int_{\xi=0}^{\xi=L}\int_{\eta=0}^{\eta=\xi}\rho(\eta)d\eta d\xi\right) \qquad (1)$$

where σ^2 is the population variance, and ρ its spatial autocorrelation function.
This variance of the mean decreases with increasing *L* but the variance of *Z* itself, at a point within *L,* increases (point properties within large volumes are more heterogeneous than the same property within a small volume). The relationship among the variances is described by Krige's relationship (Journel and Hujbregts, 1978):

$$\sigma^2_{\mathrm{o/D}} = \sigma^2_{\mathrm{o/L}} + \sigma^2_{\mathrm{L/D}} \qquad (2)$$

where $\sigma^2_{\mathrm{o/D}}$ is the variance of *Z* at a point (o) in a large volume *D*,

$\sigma^2_{\mathrm{o/L}}$ is the variance of *Z* at a point in a small volume *L* (<*D*), and

$\sigma^2_{\mathrm{L/D}}$ is the variance of the average of *Z* over *L* within *D*.

In the notation of Equation (1) $Var\left(\overline{Z}\right) = \sigma^2_{\mathrm{L/D}}$ and $\sigma^2_{\mathrm{o/D}} = \sigma^2$. The population is identified with the volume D. The variance of a point within *L* is then given by:

$$Var(Z) = \sigma^2 - Var\left(\overline{Z}\right) = \sigma^2 - \frac{2\sigma^2}{L^2}\left(\int_{\xi=0}^{\xi=L}\int_{\eta=0}^{\eta=\xi}\rho(\eta)d\eta d\xi\right) \qquad (3)$$

The integrals in the above expression can be evaluated, analytically for simple autocovariance functions, and numerically for nearly all others of interest. In particular, for the *K*-scale exponential autocovariance:

$$\rho = \sum_{\mathrm{k=1}}^{\mathrm{k=K}} f_{\mathrm{k}} e^{-\eta/\lambda_{\mathrm{k}}} \ ,$$

We have,

$$Var\left(\overline{Z}\right) = \frac{2\sigma^2}{L}\sum_{\mathrm{k=0}}^{\mathrm{k=K}} f_{\mathrm{k}}\lambda_{\mathrm{k}}\left(1-\frac{\lambda_{\mathrm{k}}}{L}\left[1-e^{-L/\lambda_{\mathrm{k}}}\right]\right) \qquad (4)$$

Equation 4 allows specification of the scales of the variability of *Z* (through the λ_k) and the fraction that each contributes to the total variance (through the f_k). These scales can in turn be related to other observations on the basis of laboratory, bed and interval scales, or even to such vagaries as micro, macro and mega. Since $1 = \sum_{\mathrm{k=1}}^{\mathrm{k=K}} f_{\mathrm{k}}$, Equation, (4) is a 2*K*- parameter model (*K*–1 f_k, σ^2, and *K* λ_k) .

Expressions for the variance of averages can be further developed for combinations of the three-scale and the stable model. Application of Equation (4) to explain the concept of Representative Elementary Volume and to understand the scaling characteristics of porosity is discussed in Lake and Srinivasan, 2004. The behavior of permeability in Fig. 1 can also be explained in this fashion as we show next.

Recall that *Z* in Equation (1) is a Gaussian scalar random variable. Hence $k = e^{Z}$ is a log-normally distributed random variable. Furthermore, assuming a uniformly layered permeability medium, the arithmetic average (expectation) of *k* denoted by k_H, is a surrogate for horizontal permeability and the harmonic average of *k*, k_V is a surrogate for vertical permeability. The direction of the scale *L* in the above equations is perpendicular to these layers.

The non-centered moments of order *j* for a log-normal distribution are given as

(Aitchison and Brown, 1976): $\lambda'_j = e^{j\mu + \frac{1}{2} j^2 \sigma^2_{o/L}}$ (5)

In Eq. 5 $\mu = E(Z) = E(\ln k) = \ln k_G$ where k_G is the geometric mean of the log-normally distributed *k*. $\sigma^2_{o/D}$ is related to $Var(\bar{Z})$ from Equation (2). Hence as the scale *L* increases, the variance of the mean $Var(\bar{Z})$ decreases (Equation 4), and the variance of a point within that length scale increases, resulting in changes in the non-centered moments of a log-normal distribution.

Equation (5) for *j* = 1 yields the arithmetic average or horizontal permeability:

$$k_H = k_G \, e^{\frac{\left(\sigma^2 - Var(\bar{Z})\right)}{2}} \qquad (6)$$

This is shown graphically in the following figure where the increase in horizontal permeability is seen with increase in averaging distance.

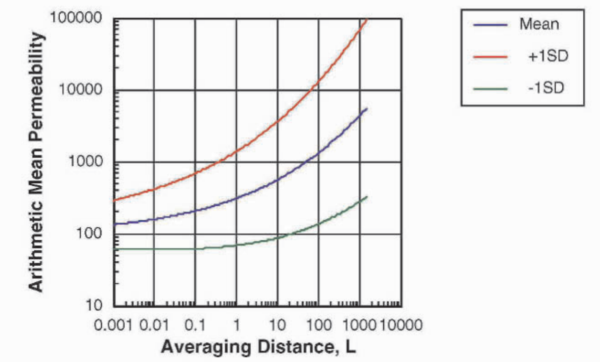


Figure 3. Calculated increase of horizontal permeability with scale calculated with Equation (6) and the one-scale stable autocorrelation model. $k_G = 100$, $\sigma^2 = 20,\ \alpha = 0.2,\ \lambda = 10000$.

Figure 3 also shows the ± 1 standard deviation of the estimate since this is available from the variance of the mean. The parameters used to generate Figure 3 are reasonable. The range parameter $\lambda = 10000$ is greater than the maximum dimension shown with $\alpha = 0.2$ being consistent with other literature (Jennings, 2000). The decrease in vertical permeability with scale can be explained similarly.

Velocity Field Scale Up

The decrease in ultimate recovery efficiency behavior in Fig. 2 and the increase in dispersivities with scale largely depend on the spatial distribution of the fluid velocities. Attempts to quantify both the dispersivity and recovery dependence on scale have implicitly assumed that the autocorrelation structure of the local fluid velocity fields is the same as that of the transport coefficients (permeability or transmissibility). This section investigates the correspondence between the permeability field and the resulting velocity field using a simple numerical flow simulation. We also evaluate the effect of vertical to horizontal permeability ratio on the spatial structure of the velocity field. The porosity is spatially constant in all the results shown below. All simulation runs were done using Eclipse(Eclipse, 2003); all semivariograms were calculated using the GSLIB (Deutsch and Journel, 1998) program *gam*.

We will simulate steady-state, single phase flow, along a reservoir cross section. The base case is a two dimensional grid with 100 grid blocks in the x direction (the direction of main flow) and 50 blocks in the z direction. Constant rate injection occurs into a well at the left edge of the cross-section (Fig. 4) and production is at a constant pressure on the right edge. Flow at three different scales are modeled using rectangular grid block sizes of 5m (base case), 10m and 20m in the x direction and sizes of 1m (base case), 2 m and 4 m in the vertical direction. These correspond to increases by factors of two and four from the base case. The injection rates are respectively doubled and quadrupled to preserve the local (cell by cell) pressure gradient/velocity ratio.

The horizontal permeability field for the base case is generated by conditional Gaussian simulation. The average (25 realizations) semivariogram is also shown in Figure 4. The difference between the semivariogram model input to the stochastic simulation and the semivariogram reproduced over the suite of realizations is because of the influence of the conditioning data. The conditioning data are along the two wells located at the edges of the cross-section. The resultant simulations all exhibit distinct stratification; properties are fairly homogeneous within the strata. This gives rise to the zonal anisotropy behaviour observed in the reproduced semivariograms. The vertical permeability k_z is assigned to be a constant fraction of the horizontal permeability k_x. This is fairly standard simulation practice the accuracy of which, however, is not known. For most generality, k_z should be simulated stochastically as are the k_x.

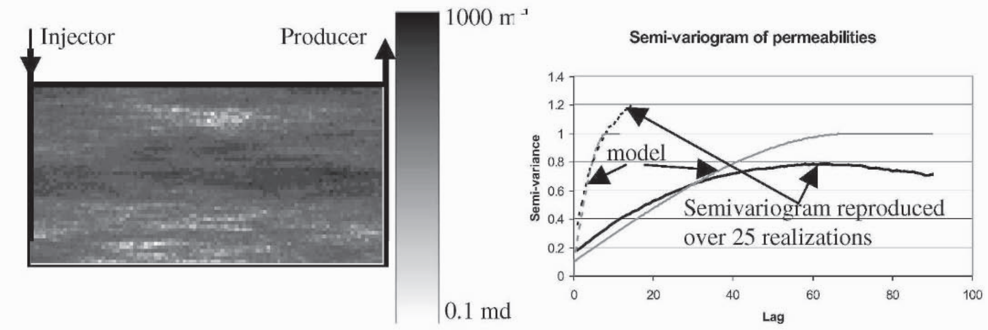


Figure 4. Fine scale x-direction permeability field and the corresponding semivariogram. The continuous lines are the semivariograms in the horizontal direction while the dotted lines are in the vertical direction.

Realizations of the scaled up permeability field are obtained by conditional Gaussian simulation using scaled up semivariograms. The simulation uses point-to-block semivariograms:

$$\bar{\gamma}(V',v) = \frac{1}{|V'|}\sum_{v'=1}^{V'}\gamma(v,v') \qquad (7)$$

and block-to-block semivariograms:

$$\bar{\gamma}(V,V') = \frac{1}{|V||V'|}\sum_{v=1}^{V}\sum_{v'=1}^{V'}\gamma(v,v') \qquad (8)$$

where V represents the scaled up block, v represents the point support assumed for the fine scale simulation (base case). Two levels of scale up were performed: i) Scale up factor of two in the x and y directions, b) Scale up factor of 4 in x and y directions. In these scaled up simulations, the original conditioning data remain as point supports and hence are reproduced only to the extent determined by the point-to-block semivariograms Eq. 7.

Figure 5 shows cross-sections of the scaled up x direction permeability field. The reduction in autocorrelation lengths for k_x in the horizontal direction can be observed; this is confirmed by semivariogram analysis. The scaled up models exhibit more variability in both the x and z directions. The sill of the horizontal semivariogram progressively increases to the normalized sill of 1.0 as the reservoir scale is increased. This is because the permeability becomes more disorganized in both the x- and z-directions as the scale is increased.

The simple scaling seems to work in that the place at which the x direction semivariance levels off is reduced by that same factor in which the grid block sizes are increased. Figure 5 suggests that the autocorrelation of the permeability is preserved on scale up Next we investigate how the velocity changes on scale up.

First, we look at the_characteristics of the x direction velocity maps as a function of the k_Z/k_X ratio. Flow was simulated on the fine scale model assuming a range of k_Z/k_X ratios. Semivariograms of the resulting horizontal and vertical velocities are computed as shown in Figure 6. As the k_Z/k_X decreases, the variability of the vertical velocities is large while that of the horizontal velocity is small. This would result in a decrease in the variance or sill of the semivariogram in the x-direction. This expected characteristic is confirmed by the semivariograms plotted in Figure 6.

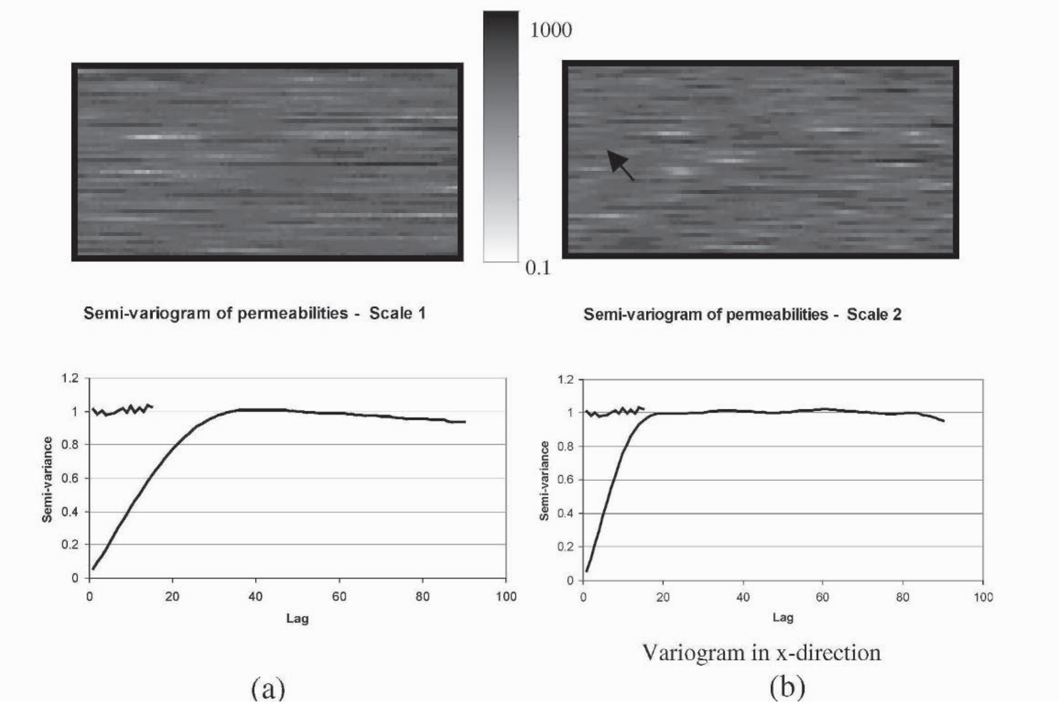


Figure 5 : Fine scale permeability model in Figure 4 scaled up: a) by a factor of 2; b) by a factor of 4. The corresponding semivariograms are also shown.

In the limiting case of zero $k_Z/k_{X=0}$ (no crossflow), all flow is horizontal through each of the layers. Though there is variation along the horizontal direction, at steady state, the horizontal velocities are the same within a layer and vertical velocities are zero. The general trend inferred from the semivariograms, is that of both velocities becoming more heterogeneous (the semivariogram levels increase) with an increase in k_Z/k_X.

The larger k_Z/k_X tend to make the flow more in vertical equilibrium (VE). VE, the state where the potential gradients in the vertical direction are zero, was shown by Arya et al. (1988) to apply to stochastically heterogeneous fields. Using this reasoning (Lake, 1989), the semivariograms for the x-direction velocity should stabilize as k_Z/k_X increases. The x direction velocity semivariograms seem to be becoming closer together as k_Z/k_X increases, but stabilization does not occur.

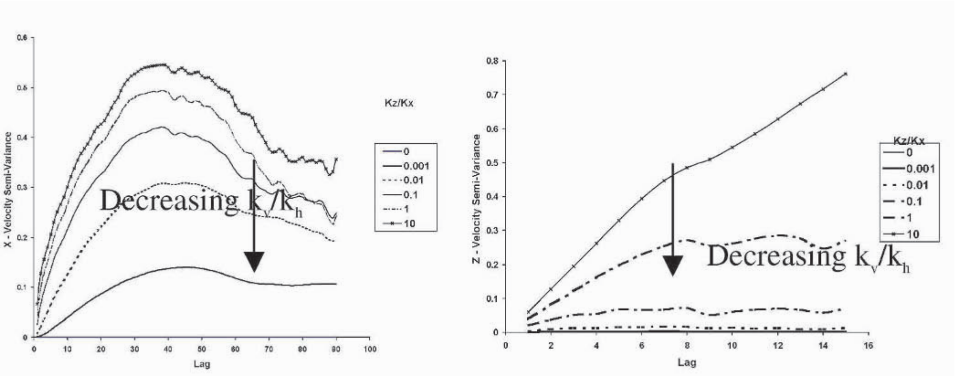


Figure 6: Single realizations of horizontal and vertical velocity semivariograms (unnormalized) corresponding to different k_Z/k_X. The vertical axis is in units of $(m/s)^2$.

Next, the velocity maps corresponding to the base case and the two scaled up cases were analyzed. k_Z/k_X =0.75 for all three cases of the flow simulation. These are shown in Figures 7. For brevity, only the velocity maps for the two extreme cases: fine scale and for a scaling factor of 4 are shown. The velocities in the horizontal and vertical directions are shown for each case.

The distinct organization of the heterogeneity into strata observed in the fine scale realization results (Fig. 3) in a well organized velocity map with a distinct streak of high velocity connecting the injector to the producer (Figure 7 (a)). At larger scales, the velocity map exhibits more variance in areas away from the producer. The layered characteristic of the permeability field in the fine scale case also causes the vertical velocities to be small in regions outside the high permeability strata. The small vertical velocities mean less sweep and consequently less recovery as in Fig. 1.

In contrast, the vertical velocities in the scaled up cases are significantly larger throughout the cross section. This suggests that the linear scaling up procedure will not reproduce the decrease in sweep in the vertical direction.

Figure 8 shows the semivariograms of the resultant x- direction velocities. At long lags, a distinct non-stationarity occurs as evidenced by the upward turn of the plots. We think this is because of the effect of the well placement and the flood direction. This behavior serves to remind us that well conditions can affect the statistics of the velocity field. We note that the x-direction permeabilities (Fig. 4) do not attain a constant plateau either.

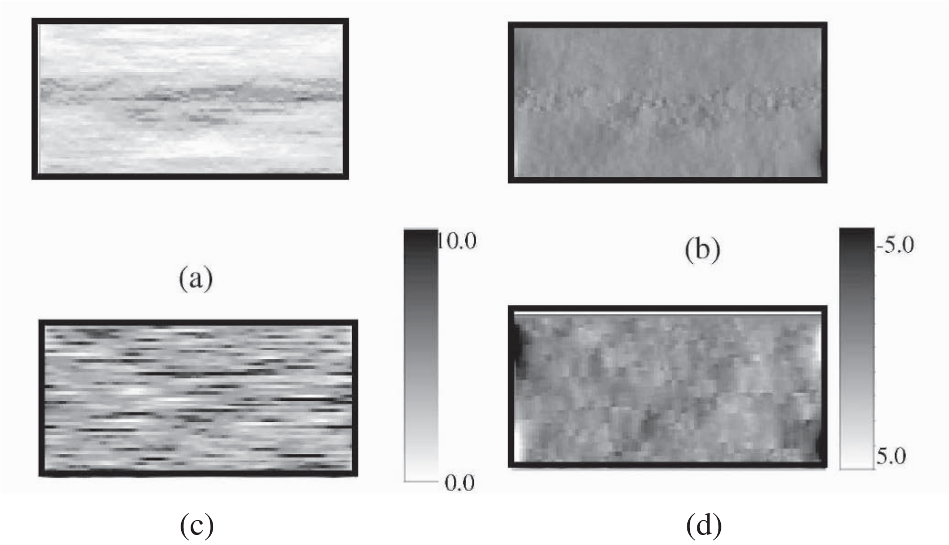


Figure 7: Spatial variations exhibited by the fluid velocity fields (m/day): a) Velocity in the x-direction for the base case, b) Velocity in the z-direction for the base case, c) Velocity in the x-direction for the reservoir model scaled up by a factor of 4, d) Velocity in the z-direction for the reservoir model scaled up by a factor of 4.

The semivariogram characteristics at intermediate lag distances do reveal subtle differences. The anisotropy ratio of the semivariograms in the x- and z-directions decreases as the reservoir scale is increased. This again indicates increased variance of the velocity field as the reservoir scale is increased. Increased variability of the scaled up permeability field causes the velocity field to exhibit more variance. The sills of the x-direction semivariograms gradually increase to the normalized value of 1.0 as the reservoir scale increases indicating departure from stratified flow conditions as the reservoir scale is increased. The range of the semivariograms decrease as the scale is increased, indicative of the dissipation of a displacement front before it reaches the producer.

This simplified numerical experiment nevertheless suggests the need for an approach to scale up properties taking into account the physics of the flow and not be limited to the simplified measures of spatial continuity.

Conclusions

The variance of the mean and how it depends on averaging scale can explain, at least qualitatively, the change in horizontal and vertical permeability with scale using reasonable autocorrelation functions and reasonable parameters for the autocorrelation functions. This is a consequence of the various means depending on variability (for non-Gaussian distributions) and the variability in turn depending on scale. Since the averages now depend on scale, it is possible that the ideas could be used to generate scale-up factors if the parameters of the autocorrelation function can be inferred from data. We briefly looked at linear scaling (multiplying permeabilities by constant factors and adjusting semivariograms) in this regard.

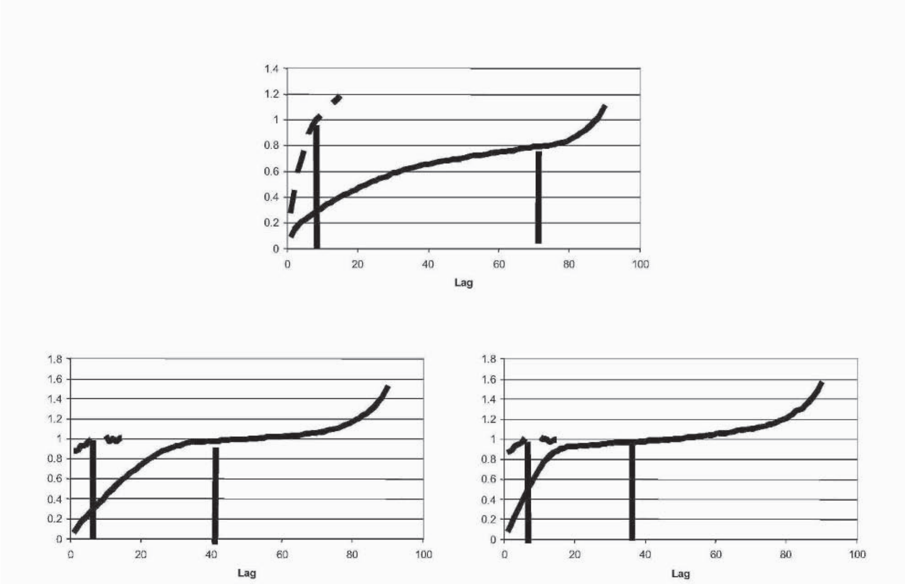


Figure 8: Semivariogram of the velocity fields corresponding to different scales of the permeability field. The vertical lines indicate the decrease in semivariogram ranges in the horizontal and vertical directions.

References

Aitchison, J., and J. A. C. Brown, *The Lognormal Distribution*, Cambridge University Press, 1976.

Arya, A., Hewett, T.A., Larson, R.G., and Lake, L.W., "Dispersion and Reservoir Heterogeneity," SPE Reservoir Engineering, Vol. 1, No. 1, February 1988, pp. 139-148.

Bear, J., *Dynamics of Fluids in Porous Media*, Amsterdam: Elsevier Scientific Publishing Co., 1972.

Datta-Gupta, A., D.W.Vasco and J.C.S. Long, "Detailed characterization of a fractured limestone formation using stochastic inverse approaches," *SPE Formation Evaluation*, 10 (3), pages 133-140, 1995.

Jennings, James W., Jr., "How much core-sample variance should a well-log reproduce," *SPE Reservoir Evaluation and Engineering*, Vol. 2, No. 5, pages 442-450, 1999.

Jennings, James W., Jr., "Spatial statistics of permeability data from carbonate outcrops of West Texas and New Mexico: Implications for improved reservoir modeling, Bureau of Economic Geology," The University of Texas at Austin, 2000.

Journel, A.G. and C. Hujbregts, *Mining Geostatistics*, Academic Press, New York City, 1978.

Lake, Larry W., Enhanced Oil Recovery, Prentice Hall, 1989.

Lake, L.W. and S. Srinivasan, "Statistical scale-up: Tools for forecasting production under uncertainty, The Journal of Petroleum Science and Engineering, 2004.

Kiraly L., "Rapport sur l'ètat actuel des connaissances dans le domaine des charactères physiques des roches karstiques," In: Burger A. and Dubertret L. (Eds.), *Hydrogeology of Karstic Terrains*, International Union of Geological Sciences, Series B 3, pages 53-67, 1975.

Mahadevan, Jagannathan, Larry W. Lake and Russell T. Johns, "Estimation of true dispersivity in field scale permeable media," Society of Petroleum Engineers Journal, Sept. 2003, pp. 272-279.

Neuman, Shlomo P., "Generalized scaling of permeabilities: Validation and effect of support scale," Geophysical Research Letters, vol. 21, no. 5, pp. 349-352, March 1, 1994.

Varela, O.J., Torres-Verdin, C and Lake, L.W., "Assessing the value of 3D seismic data in reducing uncertainty in reservoir production forecasts," SPE 77359, presented at the 2002 SPE Annual Technical Conference and Exhibition, San Antonio, Texas, September 29 - October 2.

Deutsch, C.V. and Journel, A.G., *GSLIB: Geostatistical Software Library and User's Guide*, 2nd Edition, Oxford University Press, 1998.

ECLIPSE reference manual(2003a). Schlumberger Geoquest, Geoquest Reservoir Technologies, Houston, TX, 2002.

COUPLING SEQUENTIAL-SELF CALIBRATION AND GENETIC ALGORITHMS TO INTEGRATE PRODUCTION DATA IN GEOSTATISTICAL RESERVOIR MODELING

XIAN-HUAN WEN, TINA YU and SEONG LEE
Chevrontexaco Energy Technology Company, San Ramon, CA 94583, USA

Abstract. The sequential-self calibration (SSC) method is a geostatistical-based inverse technique that allows fast integration of dynamic production data into geostatistical models. In this paper, we replace the gradient-based optimization in SSC by genetic algorithms (GA). GA, without requiring sensitivity, searches for global minimum. Although GA is computationally intensive, it provides significant flexibility to study parameters whose sensitivities are difficult to compute, e.g., master point locations. A steady-state GA is implemented under the SSC framework for searching the optimal master point locations, as well as the associated optimal perturbations that match the observed pressure, water cut and saturation data. We demonstrate that GA is easy to implement and results are robust. We examine different approaches of selecting master point locations including fixed, stratified random, and purely random methods. Results from this study demonstrate that there are not clear preferential master point locations that are best suited for matching production data for the given well pattern and for the given initial model. This is consistent with the early findings that master point locations can be randomly selected with the stratified random method yielding the best results due to its flexibility and good control for the overall model.

1 Introduction

Geostatistical reservoir models are widely used to model the heterogeneity of reservoir petrophysical properties, such as permeability and porosity. Geostatistical reservoir models must incorporate as much available, site-specific information as possible in order to reduce uncertainty in subsurface characterization, as well as in reservoir performance forecasting. Static data, such as core measurements, well logs, and seismic data, can relatively easily be integrated into geostatistical models using the traditional algorithms via conditional simulation (Deutsch and Journel, 1998). Integration of dynamic data, such as pressure, flow rate, fractional flow rate, and saturation data, is, however, a very difficult inverse problem and requires the solutions of the flow equations many times (Sun, 1994; Tarantola, 1987).

Geostatistically-based dynamic data integration has been an active area of research and a number of techniques have been reported in the literature (see Yeh, 1986 or Wen et. al., 1997 for review). The main objective is to match production data by modifying the initial geostatistical model in such a way that it preserves the underlying geostatistical

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 691–701.

features built into the initial model, such as histogram, variogram, and other soft constraints.

The Sequential Self-Calibration (SSC) method has been shown to be very efficient and robust previously for integrating dynamic production data (Gomez-Hernandez et. al., 1997; Wen et. al., 1998, 2002). The SSC uses an optimization process to modify the original reservoir model. The efficiency of the SSC method comes from the master point concept and global updating. The master points are employed to reduce the number of parameters. Perturbations at the master points are then propagated into entire model to achieve the global updating.

Current implementation of SSC uses a gradient-based optimization to compute optimal perturbation values at the master points. This requires the calculation of sensitivity coefficients that measure the changes of reservoir flow responses with respect to the change of reservoir properties. In practice, the sensitivity calculations comprise the most CPU time in the inversion. A great deal of effort has been dedicated to speed up this calculation (e.g., Vasco et. al., 1998; Wen et. al., 1998, 2003). However, the sensitivities computed by these fast methods are often inaccurate which can cause difficulty in optimization. Also, gradient-based optimizations are often trapped by local minimums for highly nonlinear problems. Furthermore, there is a strong smoothing effect when successively adding perturbation field to the initial model at each iteration, resulting in a much smoother updated model than the initial one (Wen et. al., 2002).

The main goal of this work is to implement genetic algorithms (GA) for the optimization under the SSC framework. One advantage of using GA is that both master point locations and values of reservoir properties at the master points can be considered as variables in optimization. This allows us to compare the efficiency of different selecting schemes for master point locations, as well as to investigate the possibility of any preferential master point locations for the given problem.

2 Sequential Self-Calibration (SSC) Method

The SSC method was originally developed by Gomez-Hernandez and coworkers (Gomez-Hernandez et. al., 1997). The unique features of the SSC algorithm include (1) the concept of *master point* that reduces the parameter space to be estimated in optimization, (2) the propagation procedure through *kriging* that accounts for spatial correlation of perturbations, and (3) the fast computation of sensitivity coefficients within a single flow simulation run that makes inversion feasible. The main steps of the SSC method can be summarized as follows (Wen et. al., 1998, 2002):

a) Construct initial realizations: Multiple equal-probable initial property realizations are created by conventional geostatistical methods using specific histogram and variogram consistent with the data. If static (hard and soft) data are available, they should be honored with conditional simulation. Each realization is processed one at a time with the following steps.

b) Solve the flow equations for the current model using specific boundary and well conditions to obtain flow responses.

c) Compute the objective function that measures the mismatch between the observed production data and the flow solutions. If the objective function is smaller than a preselected tolerance, this realization is considered to honor the dynamic data and we move to the next realization. Otherwise, proceed to the following steps.

i) Select a few master locations (usually 1-3 per correlation range in each direction) and solve an optimization problem to find the optimal perturbations of reservoir property at these locations.

ii) Propagate the perturbations at master locations through the entire field by kriging the computed perturbations at master points. The model is then globally updated by adding the smooth kriged perturbation field to the previous model.

iii) Loop back to step b) until convergence or enough iterations have tried.

A gradient method was previously used in step i), which requires the sensitivity coefficients (derivatives) of flow responses with respect to reservoir property changes at the selected master locations. The method for computing sensitivity coefficients of pressure has been developed previously, i.e., they are computed as part of the flow simulation run (Gomez-Hernandez, et. al., 1997; Wen et. al., 1998). The sensitivity coefficients of water cut and saturation can be computed by a fast streamline-based approach, i.e., they can be obtained by simply book-keeping streamlines in the simulation field by using the 1D analytical solution along streamlines (Wen et. al., 2003).

In this paper, we apply genetic algorithms (GA) for optimization. The advantages of using GA include (1) no need to compute sensitivities, (2) global minimum, (3) easy to implement for different type of parameters, (4) easy to honor different type of constraints built in the initial model, and (5) CPU time does not significantly increase with the number of production data. When the gradient-based optimization is used, the outer iteration, step iii), was needed to account for the non-linearity between flow and parameters since we assume linear relation during the optimization process. By using GA, we do not need the outer iteration since there is not linear assumption. Thus only one global updating is needed and the smoothing effect by adding a smooth perturbation field is reduced to minimum.

3 Genetic Algorithms (GA)

Genetic algorithms (GA) belong to the group of artificial intelligence methods. Holland first introduced and applied the principles of evolution, such as genetic inheritance and Darwinian struggle for survival, for computation (Holland, 1975). He also showed that GA is remarkable in balancing exploration and exploitation of information to perform search. Since then, GA has been applied to many optimization problems (Goldberg, 1989).

To solve an optimization problem, GA manipulates a population of individuals that is randomly initialized. Each individual represents a potential solution to the problem. The quality of each individual is evaluated by a function or a process that assigns its “fitness” to the individual. Genetic operators are applied to a population to make it evolved toward a new and "better" one. This evolutionary process is repeated as many times as desired (number of generations). From a practical standpoint, genetic algorithms are assumed to provide, in the last generation, an enhanced population where some individuals-solutions ensure the convergence of the optimization problem. Applications of GA in inverse problems have been reported by Karpouzos et. al. (2001), Romero and Carter (2001), and Yu and Lee (2002).

We use a steady-state GA (DeJong, 1975) to search for the locations and the associated permeability values of a fixed number of master points that can minimize the mismatch between the flow simulation results and observed historical production data. Steady-state GA has an overlapping population where only a portion the population is replaced at each generation. The percentage that is replaced is specified by the GA users. The selection method is the traditional roulette wheel (fitness proportionate) selection. In this method, the probability of an individual to be chosen equals to the fitness of the individual divided by the sum of the finesses of all individuals in the population. Two genetic operators are used to generate offspring: uniform crossover and Gaussian mutation. Uniform crossover picks gene values from two parents randomly to compose the offspring. Figure 1 gives an example of two offspring that are created by uniform crossover. Gaussian mutation changes a gene value to a new value based on a Gaussian distribution around the original value.

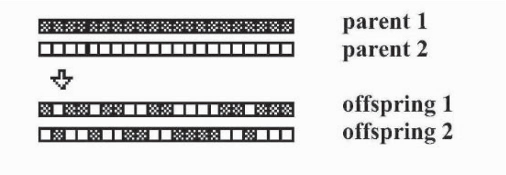


Figure 1. An example of uniform crossover.

4 Coupling SSC with GA

In this study, we only work on the permeability model. The objective function to be minimized is:

$$\boldsymbol{O} = \sum_{w_p=1}^{n_{wp}} W_{w_p}\left[\hat{p}(w_p) - p(w_p)\right]^2 + \sum_{w_f=1}^{n_{wf}} \sum_{t_f=1}^{n_{tf}} W_{w_f}\left[\hat{f}(w_f,t_f) - f(w_f,t_f)\right]^2 + \sum_{i=1}^{n_s} W_{w_s}\left[\hat{s}(i) - s(i)\right]^2 \quad (1)$$

where $\hat{p}(w_p)$ and $p(w_p)$ are the observed and simulated pressure at well w_p. $\hat{f}(w_f, t_f)$ and $f(w_f, t_f)$ are the observed and simulated water cuts at well w_f at time t_f. $\hat{s}(i)$ and $s(i)$ are the observed and simulated water saturation at cell *i* for the given time. W_{w_p} , W_{w_f} and W_{w_s} are the weights assigned to pressure, water cut, and water saturation to each well. n_{wp} and n_{wf} are the number of wells that have pressure and water cut data. n_{tf} is the number of time steps for water cut data. And n_s is the number of cells with water saturation data.

Following the SSC procedure as described above, for each initial reservoir model, we proceed the optimization process using GA. We first select a fixed number of master points and generate an initial population of initial master point locations and the associated permeability values. The master point locations are selected with three different methods: (1) fixed regular pattern, (2) stratified random (random within a regular coarse grid that covers the entire model), and (3) purely random (random within the entire model).

The permeability values at the master points are initially generated from a Gaussian function with the mean and variance consistent with the model. The constraints of the *ln(k)* values are the minimum and maximum limits. Note that if the *ln(k)* is not Gaussian, we can use a different distribution function. Also if we know the conditional pdf of each location, we can use such pdf to generate initial *ln*(*k*) values. This allows to honor different kind of constraints for the geostatistical model. Based on the generated *ln(k)* values at master point locations, as well as those at the initial model, we can compute the perturbations at the master point. We then interpolate the perturbation values at non-master point locations using kriging. An updated model is obtained by adding the perturbation field to the initial model. New fitness can be evaluated by solving flow using the updated model. If there are conditioning data that are already honored in the initial models, we include all conditioning data locations as master locations with zero perturbations. These master points are included in the GA searching process. Instead, they are simply added at the interpolation step.

With *Nm* master points (excluding the conditioning data points), the GA genome is an array of *Nm* integer and *Nm* real numbers. They are the locations and *ln(k)* values for the *Nm* master points. The steady-state GA searches for the new master point locations and the associated permeability values until the mismatch between the flow simulation results and observed historical production data is minimized. We retain the best individual (the optimal master point locations and the associated *ln(k)* values) at the end of the GA.

5 An Example

In this section, we demonstrate the applications of the coupled SSC/GA method for constructing reservoir permeability models from pressure, water cut and water saturation data using a synthetic data set. In the example, we assume porosity is known and constant as $\phi = 0.2$.

Figure 2(a) shows a 2-D geostatistical reference field (50x50 grid with cell size 80 feet x 80 feet). The model is generated using the Sequential Gaussian Simulation method (Deutsch and Journel, 1998). The *ln(k)* has Gaussian histogram with mean and variance of 6.0 and 3.0, respectively. The unit of permeability (*k*) is milli Darcy. The variogram is spherical with range of 800 feet and 160 feet in the direction of 45 degree and 135 degree, respectively. We assume an injection well (I) at the center of the model with 4 production wells (P1 to P4) at the 4 corners. The injection rate at the injection well (I) is 1600 STB/day and the production rate for the 4 production wells is 400

STB/day/well. The thickness of the reservoir is assumed constant of 100 feet. All four boundaries are no-flow boundaries. The initial pressure is constant at 3000 psi for the entire field.

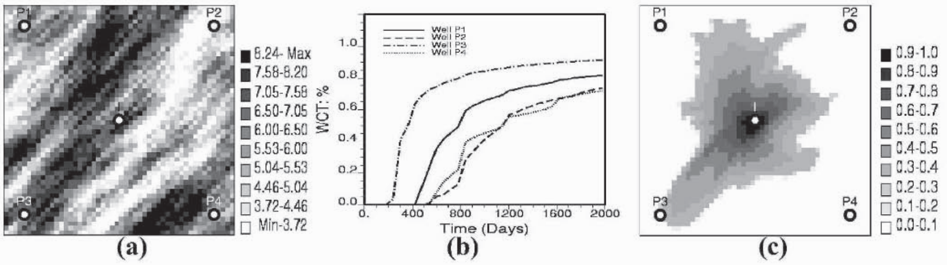


Figure 2. (a) The 2-D reference log-permeability field, (b) water cuts from the 4 production wells, and (c) water saturation distribution at 400 days.

The main features of this reference field are: (1) a high permeability zone and a low permeability zone in the middle of the field, (2) high interconnectivity between well I and well P3, (3) low interconnectivity between well I and wells P2 and P4. This reference field is considered as the true model, and our goal is to reconstruct reservoir models based on some production data that are as close to this true field as possible.

The reservoir is initially saturated with oil. Water injection and production are solved using a streamline simulator for 2000 days. Mobility ratio is 10 and standard quadratic relative permeability curves are used with zero residual saturation for oil and water. Compressibility and capillary pressure are ignored. Pressure field is updated every 400 days to account for the change of mobility during the streamline simulation. We assume the observed production data are: (1) bottom hole pressure (BHP) of each well at the end of the simulation, (2) water cut history of each production well, and (3) water saturation distribution of entire model at 400 days. These production data are supposed to mimic the practical situation of a producing field with 4D seismic survey. The “observed” BHP for I and P1-P4 are given in Table 1, the water cuts of 4 production wells and water saturation distribution at 400 days are given in Figures 1(b) and 1(c), respectively. Note that the fast water breakthrough at well P3 and late breakthrough at wells P2 and P4.

We generate multiple initial realizations using the same histogram and variogram as the reference field. These initial models are then modified to match the observed production data using the coupled SSC/GA method. We use 25 master points that are selected stratified randomly within each of the 5x5 coarse grid cells (each coarse cell represents a 10x10 fine cells).

The population size in GA is 50 and the maximum number of generations evolved is 50. The crossover rate is 90% while the mutation rate is 1%. This means that the selected two parents have 90% to be cross-over with each other to produce 2 offspring. The produced offspring (regardless if crossover has been performed or not) have 1% to be mutated. In other words, two offspring can be the results of crossover and mutation, crossover only, mutation only or identical copies. Among the population of 50 individuals, the worst 60% will be replaced with the new offspring. This is the steady-state GA explained in Section 3.

All 50 models in the last generation closely match the production data (see Figure 5). The best individual at the last generation is chosen as the final updated model. Figure 3 shows two initial permeability fields (top row) and the resulting master point locations (plus) and the perturbation fields (middle row). The final updated models are shown at the bottom of the figure. The BHPs at wells computed from the initial and updated models are given in Table 1. The water cut and water saturation matches from the initial and updated models are given in Figures 4 and 5.

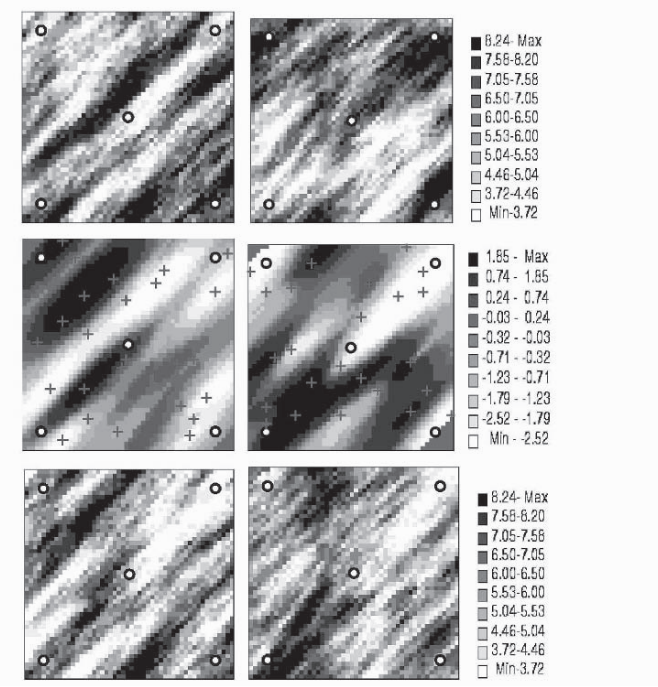


Figure 3. Two initial realizations of *ln(k)* model (top), the computed perturbation fields and master point locations (middle), and the resulting updated models (bottom).

Compared to the reference field, we can see that the spatial variation patterns in the two initial models are quite different from the reference model, resulting in significant deviations of flow responses from the “observed” production data. After inversion, the updated models display spatial variation features very similar to the reference model with flow results matching the “observed” data closely (see Figures 4 and 5). Particularly, in both models, in order to match the production data, permeabilities in the region between wells I and P3 are increased, while in the region between wells I and P2, permeabilities are reduced (see Figure 3). Based on these, we can conclude that the GA is capable of finding the optimal master point locations, as well as the associated optimal permeability values that match the production data.

Well	I	P1	P2	P3	P4
Reference	3043	2985	2468	3022	2917
Initial, #1	3135	2489	2920	2995	2974
Updated, #1	3071	2950	2422	3016	2918
Initial, #2	3037	3007	3022	2925	2872
Updated, #2	3055	2949	2505	3019	2929

Table 1. Comparison of BHPs from the two initial and updated models with the reference field

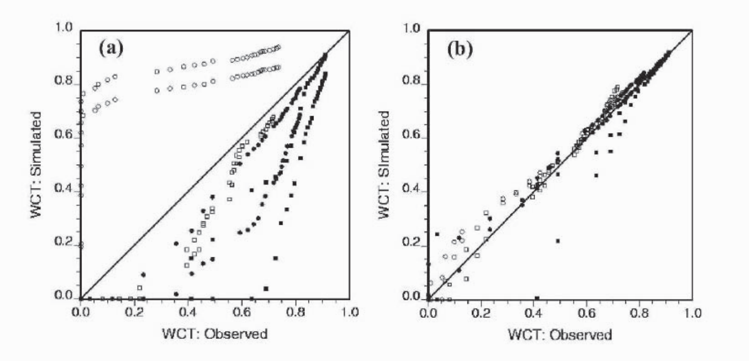


Figure 4. Scatter plots of water cuts from the two initial and updated models with respect to the observed data: (a) initial models; (b) updated models. Open circles: W1, filled circles: W2, open squares: W3, filled squares: W4.

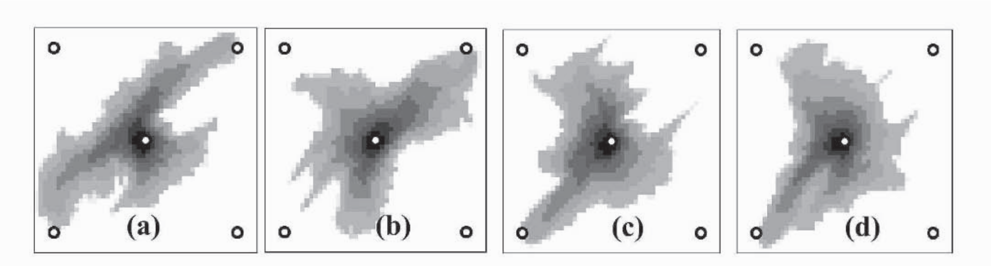


Figure 5. Water saturation distributions from the two initial and updated models: (a) initial model 1; (b) initial model 2; (c) updated model 1; (d) updated model 2. Note the reference water saturation distribution is in Figure 2(c).

Figure 6 shows the changes of objective function at each generation during the GA operation for the first model indicating the rapid reduction of objective function. The total number of function evaluation (flow simulation run) for generating one realization is about 1450.

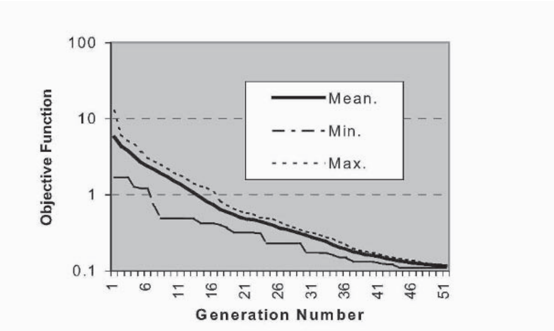


Figure 6. Variation of objective function with generation in GA.

Similar results are obtained (not shown here) by using fixed or purely random master point locations. Using fixed or pure random master point locations, however, yields the updated models with slightly larger objective function than using the stratified random method. This can be explained by (a) fixed master point locations do not provide enough flexibility on selecting best locations, (b) purely random master point locations may not provide enough overall coverage of the entire model, and (c) the stratified random method provide best compromise between the overall coverage and flexibility.

6 Discussion

Traditional pilot point method seeks the "best" pilot point locations based on sensitivity coefficients and then computes the optimal perturbations at these locations (RamaRao et. al., 1995). New "best" pilot point locations are added after each iteration. A significantly amount of CPU time is required for searching the "best" pilot point locations. The SSC method, however, uses a fixed number of randomly selected master points and computes the optimal perturbations at those locations (Wen et. al., 1998, 2002). Master point locations are updated after each several iterations during the inversion. This eliminates the time-consuming step of searching "best" locations as in the traditional pilot point method. Using the coupled SSC/GA method, one interesting issue is to investigate that, for a given well pattern or a given initial model, if there exist preferential master point locations that are superior to other locations for matching the given production data.

Figure 7 presents the total number of times that a particular cell is selected as master point from the 100 realizations using the stratified or purely random method. Clearly, there is no spatial pattern that is noticeable from these maps. Instead, they look like more or less random noise in the entire model without any structure. This demonstrates that, from a statistical point of view, there is no preferential locations that are better suited for being master point locations for the given well configurations, as long as the master points are not overly clustered in the space.

To investigate the possibility of any preferential master point locations for a given initial model, we update the same initial model 100 times using different random number seeds resulting in 100 updated models and 100 sets of master point locations. Using the two initial models as shown in Figure 3, Figure 8 shows the total number of times that a particular cell is selected as master point from the 100 runs using the stratified and purely random methods. In this case, we can see that there is slightly higher tendency that the master points are selected at areas where initial values are either too high or too low. This displays the efficiency of our method to pick up the right places to update the model. Nevertheless, this tendency is not significant indicating that there are no specific locations that are significantly better as master locations for a given initial model. Our results (not shown here) also indicate that the stratified random method provides best results in terms of the accuracy in data matching, resulting in updated reservoir models with less uncertainty compared to the fixed or purely random master point locations. From above investigation, we can conclude that master (pilot) point locations are not critical for the SSC inversion. There are no such locations that are “best” as master/pilot point locations for a given problem. Master points can be selected randomly provided that they can cover the overall model space for the given correlation structure.

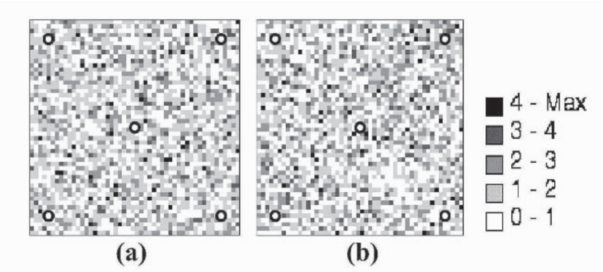


Figure 7. Total number of times that a cell is selected as master point for 100 realizations: (a) stratified random method, (b) purely random method.

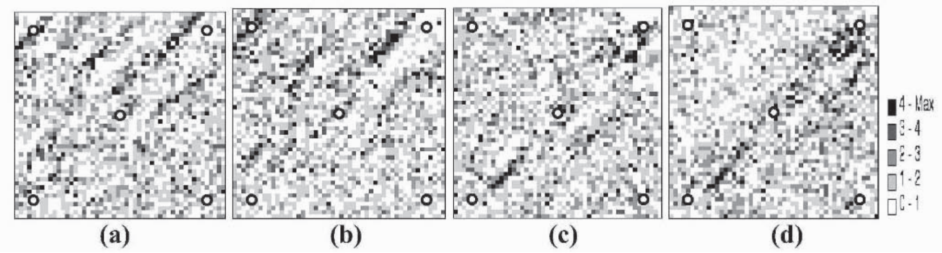


Figure 8. Total number of times that a cell is selected as master point from 100 runs using the two initial models: (a) initial model 1 and stratified random method; (b) initial model 1 and purely random method; (c) initial model 2 and stratified random method; (d)) initial model 2 and purely random method

7 Summary and Conclusions

We implemented a steady state GA under the SSC framework as an optimization process replacing the original gradient based optimization procedure. The coupled SSC/GA method is used to invert geostatistical reservoir permeability model from

This page intentionally blank

dynamic production data. The results show that the coupled SSC/GA is capable of finding optimal master point locations and the associated optimal perturbations within a reasonable number of generations. The results are accurate and robust.

GA allows us to investigate whether or not the "best" master point locations exist for a particular well pattern and for a particular initial model. We showed that there is no clear tendency with respect to where the master points should be, i.e., there are not preferential locations where the "best" master locations can be chosen. In other words, inversion results are not sensitive to what locations are chosen as master points. Master point locations can be selected randomly as long as they cover the entire model. This provides explanation to the previous studies on why randomly selected master points yielded similar results to those using "carefully" selected master points.

References

DeJong, K. A., *An Analysis of The Behavior of a Class of Genetic Adaptive Systems*, Ph.D. Thesis, University of Michigan, 1975.

Deutsch, C. V. and Journel, A. G., *GSLIB: Geostatistical Software Library and User's Guide*, 2nd edition, Oxford University Press, 1998.

Goldberg, David E. *Genetic Algorithms in Search, Optimization and Machine Learning*, Addison-Wesley Pub. Co., 1989.

Gomez-Hernandez, J. J., Sahuquillo, A. and Capilla, J. E., Stochastic Simulation of Transmissivity Fields Conditional to Both Transmissivity and Piezometric Data, 1. The theory, *Journal of Hydrology*, 203(1-4) 1997, p. 162-174.

Holland, J., *Adaptation in Natural and Artificial System*, Univ. of Mich. Press, 1975.

Karpouzos, D. K., Delay, F., Katsifarakis, K. L. and De Marsily, G., A Multipopulation Genetic Algorithm to Solve the Inverse Problem in Hydrogeology, *Water Resour. Res.*, Vol. 37 , no. 9, 2001, p. 2291-2302.

RamaRao, B.S., LaVenue, A.M., de Marsily, G. and. Marietta, M.G, Pilot Point Methodology for Automated Calibration of an Ensemble of Conditionally Simulated Transmissivity Fields, 1. Theory and Computational Experiments, *Water Resour. Res.*, 31, no. 3, 1995, p. 475-493.

Romero, C. E. and Carter, J. N., Using Genetic Algorithms for Reservoir Characterization, *Journal of Petroleum Science and engineering*, vol. 31, 2001, p. 113.

Sun N.-Z., *Inverse Problem In Groundwater Modeling*, Kluwer Academic Publishers, 1994.

Tarantola, H., *Inverse Problem Theory: Methods for Data Fitting and Model Parameter Estimation*, Elsevier, 1987.

Vasco, D. W., Yoon, S., and A. Datta-Gupta, Integrating Dynamic Data Into High-Resolution Reservoir Models Using Streamline-Based Analytical Sensitivity Coefficients*, SPE paper 49002 presented at the 1998 SPE Annual Technical Conference and Exhibition*, New Orleans, LA, Sept. 27-30, 1998.

Wen, X.-H, Deutsch, C. V. and. Cullick, A.S, A Review of Current Approaches to Integrate Flow Production Data in Geological Modeling, *In Report 10, Stanford Center for Reservoir Forecasting*, Stanford, California, 1997.

Wen, X.-H., Deutsch, C. V. and Cullick, A.S, High Resolution Reservoir Models Integrating Multiple-Well Production Data, *SPE Journal*, December, 1998, p. 344-355.

Wen, X. H., Deutsch, C. V. and Cullick, A. S., Construction of Geostatistical Aquifer Models Integrating Dynamic Flow and Tracer Data Using Inverse Technique, *J. of Hydrology*, vol. 255, 2002, p.151-168.

Wen, X. H., Deutsch, C. V. and Cullick, A. S., Inversion of Dynamic Production Data for Permeability: Fast Streamline-Based Computation of Sensitivity Coefficients of Fractional Flow Rate, *J. of Hydrology,* vol. 281, 2003, p. 296-312.

Yeh, W. W.-G., 1986. Review of parameter identification procedure in groundwater hydrology: The inverse problem. *Water Resour. Res.*, vol. 22, no. 2, 1986, p. 85-92.

Yu, T and Lee, S., Evolving Cellular Automata to Model Fluid Flow in Porous Media*, Proceedings of the 2002 NASA/DoD Conference on Evolvable Hardware*, IEEE Press, 2002, p. 210-218.

HIGH RESOLUTION GEOSTATISTICS ON COARSE UNSTRUCTURED FLOW GRIDS

GUILLAUME CAUMON[1,2], OLIVIER GROSSE[3] and
JEAN-LAURENT MALLET[1]
[1]*ENS Géologie, INPL – CRPG, Nancy, France*
[2]*Petroleum Engineering Dept., Stanford University*
[3] *Earth Decision Sciences, Nancy (France)*

Abstract. Although closely related to each other, Geostatistics and simulation of physical processes have different, often conflicting, requirements about their discretization support. While geostatistical methods can be efficiently implemented on high-resolution regular grids, process simulation calls for coarse flexible grids to minimize computational cost without loss of accuracy. Adapting geostatistical methods to such flexible grids is difficult, since unstructured neighborhood lookup is time-consuming, and cell volumes may vary significantly throughout the grid.

Instead, we propose to disconnect the representation of properties from the representation of geometry using the concept of geo-chronological space: the coarse flexible grid in present geometrical space (x, y, z) is mapped onto a high-resolution cartesian grid in geo-chronological space (u, v, t), where u and v are planar topographic coordinates at deposition time, and t is the geological time. The calculation of this 3D mapping is probably the most challenging part of the method. Here, we describe how to derive it by the extrusion of a reference stratigraphic surface, possibly discontinuous across faults. This mapping can be used to infer spatial covariance models and run geostatistical algorithms directly in geo-chronological space. The practicality of the method is demonstrated on actual reservoir data.

1 Introduction

In the petroleum and geothermal industries, the joint use of geostatistics and flow simulation has become essential to decision making. Geostatistical tools are used to generate possible petrophysical models of the subsurface that are then input to flow simulators. However complementary, geostatistics and flow simulation face a number of conflicting requirements relative to the scale, to the structure and to the geological conformity of their discretization grids.

1.1 FINE SCALE VS. COARSE SCALE

In reservoir studies, petrophysical description is typically done on fine scale grids using geostatistical algorithms, whereas flow simulation generally performs on

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 703–712.

coarse grids. Modeling petrophysical properties at a high resolution is important for representing the connectivity of high permeability values, which has a high impact on flow behavior. On the other hand, the system of flow equations on high resolution grids (with millions of cells) is too large to be solved within acceptable RAM and CPU demand on current computers. Moreover, the need for uncertainty assessment in flow response requires a large number of simulations. In practice, the time available for a particular study and the computational capability determine the resolution chosen for the flow grid.

Various methods have been defined to upscale geostatistical grids to the flow simulation scale, see Renard and de Marsily (1997), Farmer (2002), Durlofsky (2003). This paper does not explicitly address upscaling methods, although the proposed technique facilitates the application of these.

1.2 STRUCTURED VS. UNSTRUCTURED GRIDS

While the majority of geostatistical algorithms (Deutsch and Journel, 1998; Strebelle, 2002) are implemented for Cartesian grids, unstructured grids have been advocated for flow simulation (Heinemann *et al.*, 1991; Palagi and Aziz, 1991) and are becoming more and more accepted by the next generation of flow simulators.

As opposed to structured grids where connections between cells repeat periodically as in a crystal lattice, unstructured grids can define connectivity explicitly. This flexibility is interesting for flow simulation because it allows the number of cells (hence the number of flow equations) to be reduced without significant loss of accuracy (Prevost, 2003): when geological structures and well geometries are known, unstructured grids can be made to optimize the shape of grid cells and their density according to the prior flow information (Lepage, 2003).

Although the theory of geostatistics (Journel and Huijbregts, 1978; Goovaerts, 1997; Chilès and Delfiner, 1999) does not call for Cartesian grids, the implementation of a geostatistical algorithm on an unstructured grid raises a number of practical issues:

- the number of neighbors of a grid cell or node is not fixed, hence the neighborhood search is more time consuming than on structured grids.
- The cells of an unstructured grid are not necessarily aligned on the main directions of anisotropy as defined by the geological structures.
- The volume of cells varies significantly throughout the grid. This calls for accurate volume-to-volume covariance computations and carefully considering volume averaging (Deutsch *et al.*, 2002).

1.3 STATIC VS. DYNAMIC GRIDS

The geostatistical petrophysical description depends mostly on the geological structures (stratigraphy, faults), hence can be performed on one single grid if these structures are known. However, for one petrophysical model, flow can vary significantly according to the production parameters. Therefore, it several flow simulation grids can be used for a single static petrophysical model (Mlacnik *et al.*, 2004).

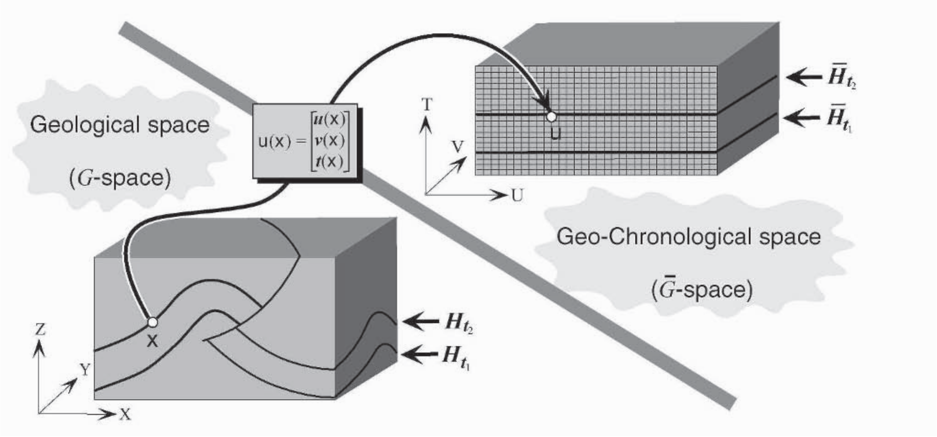


Figure 1. Any location in the geological space can be mapped onto an abstract geo-chronological space where geostatistical techniques can be applied optimally (from Mallet, 2004).

1.4 FINDING A COMPROMISE

The considerations above show the challenge of defining a gridded representation of the subsurface that can be used for both geostatistics and flow simulation. During the last decade, a significant endeavor has been set on the stratigraphic grid, a regular hexahedral grid made conform to the stratigraphy and the faults thanks to a corner point geometry (Chambers *et al.*, 1999; Mallet, 2002; Hoffman *et al.*, 2003). The stratigraphic grid is a significant improvement over the Cartesian grid, but it does not eliminate the need for upscaling. Conforming such a stratigraphic grid to complex structures (stratigraphic unconformities, complex fault networks) is difficult and often produces ill-shaped cells that may be the source of numerical instabilities when solving flow equations. Simplifications of the geological model are then required to produce acceptable flow simulation grids.

Based on the concept of geo-chronological space (Mallet, 2004), we propose to consider two distinct modeling spaces for the geostatistical and flow simulation grids (Section 2, Fig. 1): the geometrical space (G-space) defined by the geological structures and the wells is the realm of flow simulation, while geostatistical methods are applied in the depositional space ($\bar{G}$-space) where spatial correlation can be described more easily. The upscaling or downscaling of properties from one grid to another relies on a link mapping any location in the G-space to its image in the $\bar{G}$-space (Section 3).

2 Geo-Chronological space

2.1 GEOCHRONOLOGICAL SPACE AND GEOSTATISTICS

Consider a volume of the subsurface made of folded and faulted sedimentary rocks defined in a 3D geological space (G-space), where any point can be identified by a vector $\mathbf{x} = (x, y, z)$. Computing a petrophysical model for this volume using any geostatistical technique calls for computing distances between sampled and

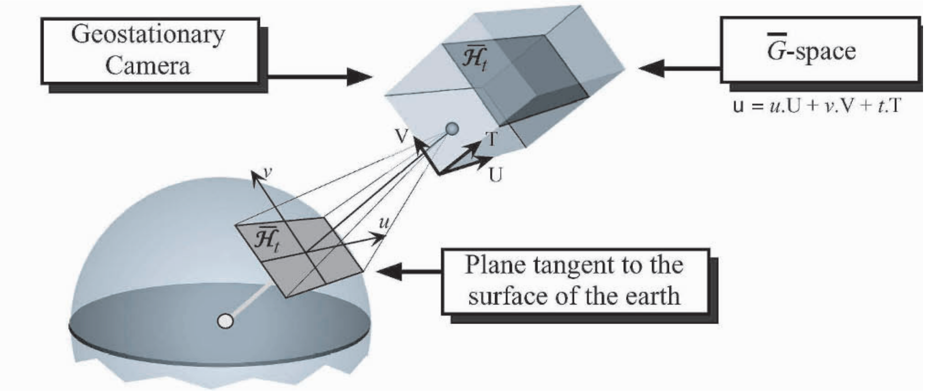

Geostationary Camera
$\overline{\mathcal{H}}_t$
$\overline{G}$-space
u = u.U + v.V + t.T
V
T
U
v
u
$\overline{\mathcal{H}}_t$
Plane tangent to the surface of the earth

Figure 2. The Geo-chronological space seen as a stack of pictures taken from a geostationary camera (from Mallet, 2004).

unsampled locations in this volume. Using Euclidean distances in the G-space is obviously not the best solution because the continuity of petrophysical properties is better along the stratigraphy. It is more appropriate to define a curvilinear coordinate system (u, v, t) aligned on the stratigraphy, then to compute distances using these coordinates (Fig. 1).

The geo-chronological model proposed by Mallet (2004) aims at defining such a coordinate system through a function that maps any location $\mathbf{x} = (x, y, z)$ of the G-space to an image $\mathbf{u}(\mathbf{x}) = (u, v, t)$ in geo-chronological space (geochron space, or $\bar{G}$-space) where spatial correlations can be described more easily.

Conceptually, the $\bar{G}$-space can be seen as a stack of aerial snapshots of the domain of study as would be taken from a geostationary camera over successive geological times (Fig. 2). In this space, any location is described by a vector $\mathbf{u} = (u, v, t)$. The horizontal (u, v) coordinates, denoted as paleo-geographic coordinates, can be used to describe the petrophysics at any location on the surface of the earth at a given geological time t. Within a horizontal plane, the sedimentological continuity is better understood because no tectonic event has yet deformed or faulted the sedimentary rocks; analogs required by multiple-point geostatistics (Caers, 2001; Strebelle, 2002; Arpat and Caers, 2004) can be obtained from present observations of the Earth surface. The geo-chronological space is thus more adapted to perform geostatistics than the geological space.

The idea of transforming the geological units to a space where they can be understood more easily was initially proposed by Wheeler (1958). The geochron model is a mathematical formulation of this time-stratigraphy concept, and relies on the interpretation of time-significant surfaces such as maximum flooding surfaces and transgression surfaces from the seismic data (Vail *et al.*, 1977).

2.2 MAPPING THE GEO-CHRONOLOGICAL SPACE TO THE PHYSICAL SPACE

The geo-chronological space may seem of little interest to the issue of flow prediction: flow simulation makes full sense only in the physical space, where wells and

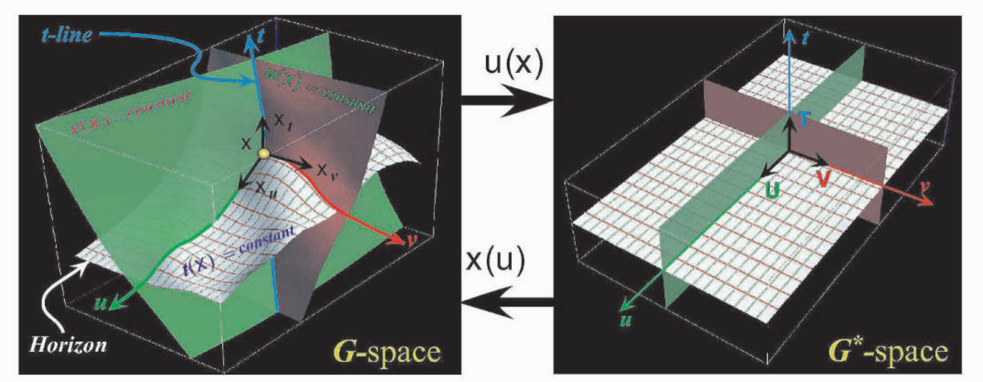


Figure 3. The orthonormal coordinate system in the $\bar{G}$-space maps onto a curvilinear coordinate system in the G-space.

volumes can be modeled. We must therefore define how to transfer petrophysical properties from the $\bar{G}$-space to the G-space, where the flow simulation grid(s) is (are) embedded.

However, defining a one-to-one mapping between the G-space and the $\bar{G}$-space is not always possible, due to erosion processes. Consider then the subset $\bar{G}_0$ of the $\bar{G}$-space which has never been eroded, defined by:

$$\mathbf{u} = (u, v, t) \in \bar{G}_0 \quad \Leftrightarrow \quad \exists \mathbf{x} = (x, y, z) \in G \text{ such as } \mathbf{u}(x, y, z) = (u, v, t)$$

$\bar{G}_0$ is referred to as the parametric domain of the geological space. For any location (u, v, t) of $\bar{G}_0$, it is now possible to define the inverse function of $\mathbf{u}(x, y, z)$ noted $\mathbf{x}(u, v, t)$. This function $\mathbf{x}(u, v, t)$ induces a curvilinear coordinate system (u, v, t) in the G-space defined by u-lines, v-lines and t-lines depicted in Figure 3. Note that t-lines may differ from the surface normal if tectonic deformations included a flexural slip component. Computing u-, v- and t-lines in the $\bar{G}$-space amounts to defining the functions $\mathbf{u}(x, y, z)$ and $\mathbf{x}(u, v, t)$, that is, mapping any location in the G-space to its image in the $\bar{G}_0$ domain and conversely. Section 3 will present a practical way to approximate such lines.

2.3 GEO-CHRONOLOGICAL MODEL, GEOSTATISTICS AND FLOW SIMULATION

How can the geo-chronological model help in addressing the conflicting requirements between geostatistical grids and flow simulation grids? Instead of striving to obtain an ideal (and probably unattainable) grid that would suit the practical needs of both geostatistics and flow simulation methods, we suggest to use two distinct grids which can each be optimized for the algorithm retained.

The geostatistical grid $\bar{\mathcal{G}}$ is defined as a high-resolution Cartesian grid in geochronological space. Most existing algorithms can be applied on such a grid (Deutsch and Journel, 1998; Strebelle, 2002); a conditioning datum at location

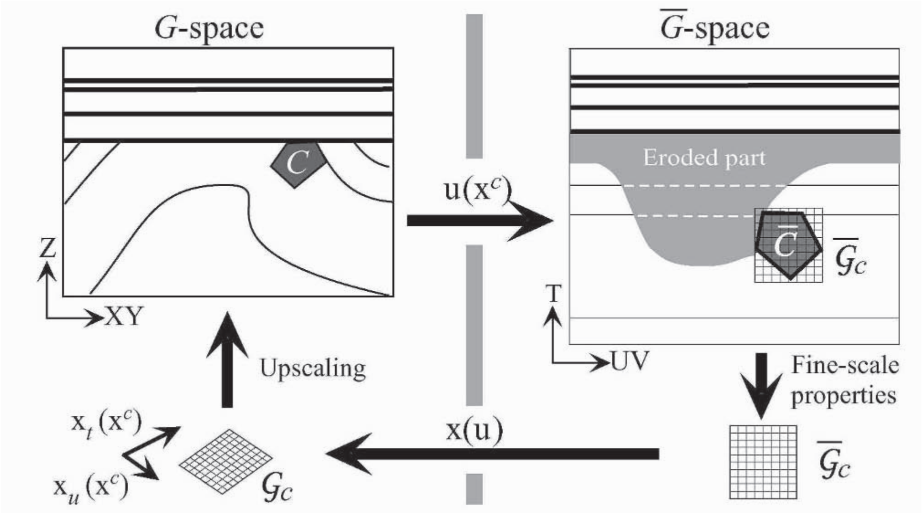


Figure 4. Computing equivalent property on a coarse flow grid in G-space by reading and upscaling the high-resolution geostatistical description in $\bar{G}$-space (from Mallet, 2004).

(x, y, z) just needs be transferred to the grid $\bar{\mathcal{G}}$ using the mapping function $\mathbf{u}(x, y, z)$.

The flow simulation grid $\mathcal{G}$ is defined in the present geological space. It is made conform to the horizons and faults that represent significant petrophysical discontinuities. For any cell c of the flow grid $\mathcal{G}$, fine-scale petrophysical properties can be read in a subset $\bar{\mathcal{G}}_c$ of the geostatistical grid $\bar{\mathcal{G}}$. The equivalent coarse-scale properties can then be computed as suggested by Figure 4 by the appropriate upscaling approach (Renard and de Marsily, 1997; Farmer, 2002; Durlofsky, 2003). Note that upscaling results will be consistent even if the subset $\bar{\mathcal{G}}_c$ is slightly larger than the image $\bar{c}$ of the cell c, since the geostatistical grid does not display petrophysical discontinuities induced by faults and erosion.

3 Defining the geochronological transform on a flow simulation grid

In this section, we present a practical way to approximate the geo-chronological parameterization while creating a polyhedral flow simulation grid conforming to geological structures and to wells.

3.1 PALEO-GEOGRAPHIC COORDINATES (U, V)

Consider a horizon $\mathcal{H}_t$ deposited at a geological time t, possibly folded and faulted. The image $\bar{\mathcal{H}}_t$ in the $\bar{G}$-space of this horizon $\mathcal{H}_t$ is by definition a plane, corresponding to a virtual aerial picture of the domain of study at time t (Fig. 2).

There is an infinite number of ways to map the horizon $\mathcal{H}_t$ onto its planar and continuous image $\bar{\mathcal{H}}_t$. We suggest that deformations should be minimal between the G-space and the $\bar{G}$-space. For this, consider the partial derivatives of the parametric function $\mathbf{x}(u, v, t)$ with respect to u, v and t (Figure 3):

$$\begin{array}{rcl} \mathbf{x}_u(\mathbf{x}) & = & \left.\frac{\partial \mathbf{x}(u,v(\mathbf{x}),t(\mathbf{x}))}{\partial u}\right|_{u=u(\mathbf{x})} \\ \mathbf{x}_v(\mathbf{x}) & = & \left.\frac{\partial \mathbf{x}(u(\mathbf{x}),v,t(\mathbf{x}))}{\partial v}\right|_{v=v(\mathbf{x})} \\ \mathbf{x}_t(\mathbf{x}) & = & \left.\frac{\partial \mathbf{x}(u(\mathbf{x}),v(\mathbf{x}),t)}{\partial t}\right|_{t=t(\mathbf{x})} \end{array}$$

It can be shown (Mallet, 2004) that minimizing deformations amounts to honoring the three following constraints as much as possible:

$$\begin{array}{rl} \|\mathbf{x}_u(\mathbf{x})\| = \|\mathbf{x}_v(\mathbf{x})\| = 1 & \text{(conservation of distances)} \\ \mathbf{x}_u(\mathbf{x}) \cdot \mathbf{x}_v(\mathbf{x}) = 0 & \text{(conservation of angles)} \end{array} \tag{1}$$

3.2 ADDING THE TIME COORDINATE T

Computing the time coordinate t does not require knowledge of the actual geological age of the horizons, but only of their relative age: arbitrary times can be used if they are sorted in increasing order from the oldest to the youngest terrains. Between any two horizons $\mathcal{H}_{t_1}$ and $\mathcal{H}_{t_2}$ deposited respectively at times t_1 and t_2, we propose to approximate the t-lines by a field of vectors made tangent to the faults and the boundary of the domain of study. This field can be interpolated away from faults using Discrete Smooth Interpolation (Mallet, 2002, p. 379). This approximation of the Geochron model is valid only if we can assume that faults do not cross t-lines; distortions of the paleo-geographic coordinates may be otherwise generated by the method.

A reference horizon $\mathcal{H}_{t_i}$ is then selected and provided with a paleo-geographic coordinate system that honors the constraints (1). This can be done using a surface parameterization algorithm (Lévy and Mallet, 1998; Mallet, 2002; Lévy *et al.*, 2002). By definition, the $(u(\mathbf{x}), v(\mathbf{x}))$ coordinates are constant along a given t-line, and can thus be propagated between $\mathcal{H}_{t_1}$ and $\mathcal{H}_{t_2}$ along the t-vectors.

The geological time $t(\mathbf{x})$ can be interpolated along this vector field, accounting for both high-resolution well data between horizons $\mathcal{H}_{t_1}$ and $\mathcal{H}_{t_2}$ and expert information on the stratigraphic style (Caumon, 2003, chap. 1).

This approach provides a reasonable approximation of the t-lines for geological models with subvertical or listric faults. At any location $\mathbf{x} = (x, y, z)$ in the G-space, it is then possible to obtain by interpolation the $(u(\mathbf{x}), v(\mathbf{x}), t(\mathbf{x}))$ coordinates in the $\bar{G}$-space.

3.3 CREATION OF THE FLOW GRID

The geo-chronological parameterization above can also be used to create a polyhedral flow grid conforming to the geological structures and to subvertical wells. The mesh of the reference horizon $\mathcal{H}_{t_i}$ can be modified in the (u, v) parametric space, to account for heterogeneities and well configuration; the grid is then created by extrusion along the t-lines, (Fig. 5-C, D).

When some layers in the G-space pinch out due to unconformities, 0-volume cells need not be created. This simply means that some petrophysical properties in the $\bar{G}$-space will not be used in the flow grid, at locations which have been eroded.

3.4 APPLICATION TO A PETROLEUM RESERVOIR

The proposed methodology was applied to a North Sea petroleum reservoir depicted in Figure 5. The hanging wall of the reservoir was parameterized, and a field of t-vectors was computed and made tangent to the faults (Fig. 5-A). Using these, nine wells could be transferred into the geo-chronological space. There, two rock types were simulated with snesim (Strebelle, 2002) using a fluvial-type reservoir training image. The property grid was then populated with porosity and permeability using Sequential Gaussian Simulation within each rock type. The results of one realization of porosity are shown in full resolution in the Geochron space (Fig. 5-B) and in the geological space (Fig. 5-C). The upscaled porosity model as obtained by simple arithmetic averaging is displayed in Figure 5-D.

4 Conclusion

The proposed method ties a high resolution geostatistical grid defined in geochronological space to one or several flow simulation grids defined in geological space. The mapping proposed makes it possible to apply upscaling and downscaling from one grid to another. The method is similar in spirit to the concept of stratigraphic grids, but presents several improvements:

- the flow grid need not be made of hexahedral cells. This increased flexibility make modeling of faults easier while reducing the number of cells. Radial grid geometries can be defined around vertical wells and flow-based gridding is also possible.
- Because the petrophysical model is computed before creating the flow grid, the density of cells can be made dependent on the petrophysical heterogeneities.
- If needed, the flow grid can be rebuilt without having to maintain the definition of the petrophysical model.
- For a given amount of RAM, the property grid can have a higher resolution than a faulted stratigraphic grid since its geometry and connectivities are not stored explicitly.
- Existing geostatistical algorithms implemented for Cartesian grids can be used directly to create the petrophysical model, without caring about the type of flow grid retained.

Due to the extrusion process proposed to build the flow grid, complex fault networks cannot be handled in a satisfactory manner; an alternative approach based on a 3D parameterization of a tetrahedralized mesh is currently being investigated (Moyen *et al.*, 2004).

Another possible improvement of the method would be to model petrophysical changes after deposition, due for instance to compaction and fracturation. Such

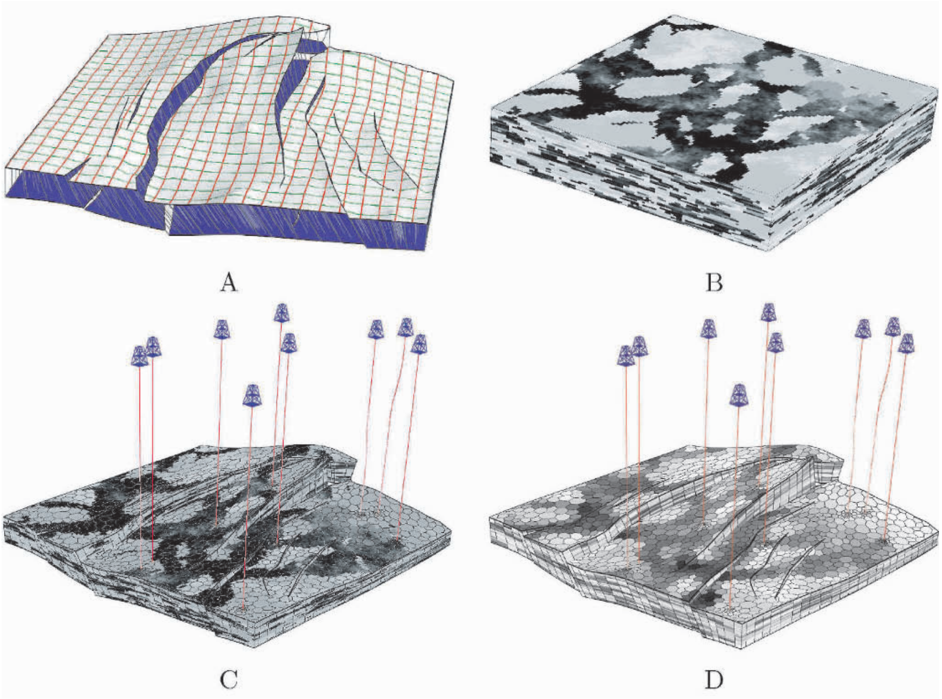


Figure 5. Paleo-geographic coordinates and approximation of t-lines by a vector field (A) on a North Sea reservoir are used to transfer conditioning data to the geochronological space, where simulations can be run (B). This high-resolution petrophysical model can be displayed (C) and upscaled (D) to the resolution of the polyhedral flow grid.

studies could benefit from the Geochron model, which can produce estimates of the strain.

Acknowledgements

We would like to thank the sponsors of the Gocad Research Consortium and particularly Earth Decision Sciences for supporting this research, and an anonymous reviewer of this paper for his suggestions. We also would like to acknowledge Stéphane Conreaux from Earth Decision Sciences and Bruno Lévy from INRIA for their pioneering work on this topic.

References

B. G. Arpat and J. Caers. A multiple-scale, pattern-based approach to sequential simulation. In O. Leuangthong and C. V. Deutsch, editors, *Geostatistics Banff, Proc. of the seventh International Geostatistics Congress*. Kluwer, Dordrecht, 2004.

J. Caers. Geostatistical reservoir modeling using statistical pattern recognition. *Journal of Petroleum Science and Engineering*, 29(3):177–188, 2001.

G. Caumon. *Représentation, visualisation et modification de modèles volumiques pour les géosciences*. PhD thesis, INPL, Nancy, France, March 2003. 150 p.

K. T. Chambers, D. R. DeBaun, L. J. Durlofsky, I. J. Taggart, A. Bernath, A. Y. Shen, H. A. Legarre, and D. J. Goggin. Geologic modeling, upscaling and simulation of faulted reservoirs using complex, faulted stratigraphic grids. In *Proc. SPE Reservoir Simulation Symposium (SPE 51889), Houston, TX*, 1999.

J.-P. Chilès and P. Delfiner. *Geostatistics: Modeling Spatial Uncertainty*. Series in Probability and Statistics. John Wiley and Sons, 1999. 696 p.

C. V. Deutsch and A. G. Journel. *GSLIB: Geostatistical Software Library and user's guide*. Applied Geostatistics. Oxford University Press, New York, NY, 2nd edition, 1998. 384 p.

C. V. Deutsch, T. T. Tran, and M. J. Pyrcz. Geostatistical assignment of reservoir properties on unstructured grids. In *SPE Annual Technical Conference and Exhibition (SPE 77427)*, 2002.

L. J. Durlofsky. Upscaling of geocellular models for reservoir flow simulation: a review of recent progress. In *Proc. 7th International Forum on Reservoir Simulation, Bühl/Baden-Baden, Germany*, June 2003.

C. L. Farmer. Upscaling: a review. *International Journal for Numerical Methods in Fluids*, 40(1-2):63–78, 2002.

P. Goovaerts. *Geostatistics for natural resources evaluation*. Applied Geostatistics. Oxford University Press, New York, NY, 1997. 483 p.

Z. E. Heinemann, C. W. Brand, M. Munkan, and Y. Chen. Modeling reservoir geometry with irregular grids. *SPE Reservoir Engineering (SPE 18412)*, pages 225–232, May 1991.

K. S. Hoffman, J. W. Neave, and R. T. Klein. Streamlining the workflow from structure model to reservoir grid. In *Proc. Annual Conference (SPE 66384)*, 2003. 7p.

A. G. Journel and C. Huijbregts. *Mining Geostatistics*. Academic Press, NY; Blackburn Press, NJ (reprint, 2004), 1978.

F. Lepage. *Génération de maillages tridimentionnels pour la simulation des phénomènes physiques en géosciences*. PhD thesis, INPL, Nancy, France, 2003.

B. Lévy and J.-L. Mallet. Non-distorted texture mapping for sheared triangulated meshes. In *Computer Graphics (Proc. Siggraph.)*, pages 343–352. ACM Press, New York, NY, July 1998.

B. Lévy, S. Petitjean, N. Ray, and J. Maillot. Least square conformal maps for automatic texture generation. *ACM Transactions on Graphics (Proc. Siggraph)*, 21(3):362–371, 2002.

J.-L. Mallet. *Geomodeling*. Applied Geostatistics. Oxford University Press, New York, NY, 2002. 624 p.

J.-L. Mallet. Space-time mathematical framework for sedimentary geology. *Mathematical geology*, 36(1):1–32, 2004.

M. Mlacnik, L. J. Durlofsky, and Z. E. Heinemann. Dynamic flow-based PEBI grids for reservoir simulation. In *Proc. Annual Conference (SPE 22889)*. SPE, 2004.

R. Moyen, J.-L. Mallet, T. Frank, B. Leflon, and J.-J. Royer. 3D-parameterization of the 3D geological space - the GeoChron model. In *Proc. European Conference on the Mathematics of Oil Recovery (ECMOR IX)*, 2004.

C. Palagi and K. Aziz. Use of voronoi grids in reservoir simulation. In *Proc. Annual Conference (SPE 22889)*. SPE, October 1991.

M. Prevost. *Accurate Coarse Reservoir Modeling Using Unstructured Grids, Flow-Based Upscaling and Streamline Simulation*. PhD thesis, Stanford University, 2003.

P. Renard and G. de Marsily. Calculating equivalent permeability: a review. *Advances in Water Resources*, 20(5-6):253–278, 1997.

S. Strebelle. Conditional simulation of complex geological structures using multiple-point statistics. *Math. Geol.*, 34(1), 2002.

P. R. Vail, R. M. Mitchum, R. G. Todd, J. M. Windmier, S. Thompson, J. B. Sangree, J. N. Bubb, and W. G. Hatledid. Seismic stratigraphy and global changes of sea level. In C. E. Payton, editor, *Seismic Stratigraphy – Applications to hydrocarbon exploration*, volume 26, pages 49–212. AAPG Memoir, 1977.

H. F. Wheeler. Time-stratigraphy. *Bull. of the AAPG*, 42(5):1047–1063, 1958.

ASSESSMENT OF UNCERTAINTY IN RESERVOIR PRODUCTION FORECASTS USING UPSCALED FLOW MODELS

OLE PETTER LØDØEN[1], HENNING OMRE[1],
LOUIS J. DURLOFSKY[2,3] and YUGUANG CHEN[2]
[1] *Department of Mathematical Sciences, Norwegian University of Science and Technology, N-7491 Trondheim, Norway*
[2] *Department of Petroleum Engineering, Stanford University, Stanford, CA 94305-2220, USA*
[3] *ChevronTexaco ETC, San Ramon, CA 94583-0719, USA*

Abstract. Production forecasts for a given reservoir, with associated uncertainties, must be based on a stochastic model of the reservoir coupled with a fluid flow simulator. These flow simulations may be computationally prohibitive if they are performed on the fine scale geostatistical model. To accelerate these computations, coarsened (or upscaled) reservoir models are generally simulated. The recovery predictions generated using the upscaled models will inevitably differ from those of the underlying fine scale models and may display severe biases and erroneous uncertainties. In this paper, a model that accounts for these biases and changes in uncertainties is described. Issues related to the selection of fine scale calibration runs and the performance of the method using different upscaling procedures are considered. Two test cases involving water injection are presented. These examples demonstrate the large errors that can result from standard upscaling techniques, the ability of the error modeling procedure to nonetheless provide accurate predictions (following appropriate calibration), and the benefits of using a more accurate upscaling procedure.

1 Introduction

Efficient reservoir management requires reliable forecasts of production for a given recovery strategy. It is also important to quantify the uncertainty associated with these forecasts, which can be accomplished through repeated sampling from a stochastic model. In this case, this sampling requires performing a fluid flow simulation. However, because geostatistical descriptions may contain 10^7-10^8 cells, performing direct flow simulations on these models may be computationally prohibitive in practical settings. These calculations must therefore be accelerated significantly to allow for the repeated sampling. This can be accomplished through upscaling of the geostatistical model or by simplifying the physics involved in

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 713–722.

the flow process. Either of these approximations will increase the computational efficiency, but they are also known to change the error structures and introduce biases in the production forecasts.

Several previous authors discussed the calibration of complex computer models and variance correction when using approximate models (Kennedy and O'Hagan, 2001; Craig et al., 2001). None of these studies, however, formalized the modeling of the bias introduced by the approximate simulations. Since bias in the forecasts of cumulative oil production can potentially have a substantial impact on predictions and thus development decisions, this should be included as part of the study. In this work, we extend and illustrate a recently developed technique for modeling the error introduced by upscaled reservoir descriptions (Omre and Lødøen, 2004). We present a procedure for determining which fine scale models to simulate (as required for the calibration of the error model) and assess the performance of the overall procedure with various upscaling techniques. The approach is applied for different flow scenarios in complex channelized systems. The benefits derived from using more accurate upscaling techniques are demonstrated.

This paper proceeds as follows. First, a statistical error model that accounts for the effects of using approximate flow simulators is presented. The methodology is then demonstrated and verified on a highly heterogeneous, channelized 2D reservoir using different upscaling algorithms and flow simulation scenarios. We show that the use of accurate upscaling procedures reduces the number of fine scale calibration runs required, resulting in potential computational savings.

2 Stochastic Model

The stochastic model is explained in detail in Omre and Lødøen (2004) and will only be discussed briefly here. The stochastic reservoir variable is termed R, and includes all the variables necessary for evaluation of fluid flow. Assume that a prior probability density function (pdf) can be assigned to R, $f(r)$, either through conventional variogram based geostatistics (Hegstad and Omre, 2001) or through multiple-point geostatistics (Strebelle, 2002). In either case the reservoir variable can be conditioned to seismic data and well observations, meaning that samples can be efficiently produced. To keep the notation simpler, data conditioning is not included in the notation here. The stochastic production variable is termed P, and contains all the information requested as output from a forward flow simulation. Hence, R will be a spatial variable, covering all the grid blocks in the reservoir domain, and P will be a time-series covering the time domain.

A fluid flow simulator links the reservoir variable R to the production variable P through $P = \omega(R)$. The chosen flow scenario must also be specified as input to the flow simulator, but this is not included in the notation here. Assuming that the flow simulator is "perfect," the conditional pdf for the production variable given the reservoir variable is fully defined through the flow simulation by $f(p|r) = \omega(r)$. Through the product rule for dependent variables, the joint pdf of interest becomes

$$f(p, r) = f(p|r)f(r). \tag{1}$$

This model is not analytically tractable, however, primarily due to the fact that the relationship between static (e.g., porosity and permeability) and dynamic (flow response) properties is nonlinear. Furthermore, the flow simulation may be computationally prohibitive if the fine scale model contains a high degree of detail (e.g., $O(10^7 - 10^8)$ cells). This prohibits direct application of algorithms that perform flow simulations on the fine scale.

To enable sampling (by reducing the computational cost), an approximate fluid flow simulator ω_* is introduced. This is most often accomplished through the use of some type of upscaling; i.e., solution of the flow equations on a coarser mesh. The approximate production variable P_* is easily obtained through $P_* = \omega_*(R)$ due to the fact that the approximate flow simulator has a low computational cost. There is no stochasticity associated with the flow simulation; i.e., the same input will always give exactly the same output. The joint pdf of the production variable, the approximate production variable, and the reservoir variable is given by

$$f(p, p_*, r) = f(p|p_*, r)f(p_*|r)f(r). \tag{2}$$

The most computationally demanding aspect of Eq. 2 is the evaluation of the pdf $f(p|p_*, r)$, which involves a flow simulation on the fine scale. If the approximate production variable captures the important features of the fine scale production variable, $f(p, p_*, r)$ can be approximated by $f_*(p, p_*, r)$ simply by modeling $f(p|p_*, r)$ with $f(p|p_*)$ as described below. By integrating over all possible realizations of the approximate production variable and all possible realizations of the reservoir variable, the following approximation is found:

$$f_*(p) = \int_{\Omega_{P_*}} \int_{\Omega_R} f(p|p_*)f(p_*|r)f(r)drdp_*, \tag{3}$$

where Ω_{P_*} and Ω_R refer to the spaces of all possible realizations of the approximate production variable and the reservoir variable, respectively. This model is still not analytically tractable due to the nonlinear relation between the reservoir properties and the coarse scale production variable. However, the pdf can be efficiently sampled if appropriate representations of $f(p|p_*)$ and $f(r)$ can be found.

An estimate of $f(r)$ can readily be found through Monte Carlo integration. The estimation of $f(p|p_*)$ is, however, not trivial. This can potentially be accomplished in a variety of ways, but an empirical, statistical approach is chosen here. The idea is to use a low number n_{fine} of calibration runs, which are performed on both the fine and coarse scales. These simulations are used to estimate a set of parameters that defines the pdf $f(p|p_*)$. This entails selecting one realization r' of the reservoir variable and performing two forward simulations to obtain $p'_* = \omega_*(r')$ and $p' = \omega(r')$. By repeating this process for n_{fine} different realizations of the reservoir variable, n_{fine} pairs of flow simulation results are obtained. This will be the calibration set. The low number of observations in the calibration set suggests a parametric estimation approach, but in principle any method that provides an estimate of $f(p|p_*)$ can be used. Performing the estimation in a multivariate linear regression setting, the estimate of $f(p|p_*)$ will become:

$$\hat{f}(p|p_*) \sim Gauss(\hat{A}_* p_*, \hat{\Sigma}_*), \tag{4}$$

where $\hat{A}_*$ and $\hat{\Sigma}_*$ are the estimates of the regression coefficient matrix and covariance matrix, respectively. For upscaling algorithms of reasonable accuracy, the error will be relatively small and this linearity assumption should be valid. For less accurate upscaling techniques, this assumption is more approximate. However, as will be illustrated in the examples below, the leading error is still well represented through a linear relationship of the form of Eq. 4.

The estimates of A_* and Σ_* may depend strongly on the choice of calibration set. This is a significant issue as there is initially no information regarding the fine scale forward simulations. Because the fine scale flow simulations are extremely time consuming, there is a need to efficiently select the realizations for the calibration set. It is also beneficial to accomplish this calibration using a minimum number of fine scale simulations.

Our calibration procedure is as follows. By treating each production variable at each point in time independently, the covariance matrix will become a diagonal matrix. This means that linear regressions must be performed for each production variable at selected points in time and that all the linear regressions will have associated prediction intervals. This allows a minimum variance criterion to be defined for the determination of the calibration runs. The idea is to minimize the size of all the confidence intervals by minimizing the trace of the covariance matrix. The choice of calibration set should not influence the estimation of the variances in the residual terms significantly, provided that the "true" error model is approximately linear (i.e., Eq. 4 holds). Hence, the selection of the calibration set can be performed based solely on the coarse scale simulation runs. Note that this is not in general a globally optimal choice of calibration set, since guaranteeing this would require knowledge of all the fine scale simulations in advance. Note also that decisions regarding the number of realizations to include in the minimum variance calibration set is presently not an automated part of the selection procedure.

By performing the calibration runs to estimate $f(p|p_*)$, and drawing a set of samples $(p_*^i, r^i); i \in \{1, ..., n\}$ from $f(p_*, r)$, an approximate forecast of the production from the reservoir can be found through:

$$\hat{f}_*(p) = \sum_i \hat{f}(p|p_*^i)\hat{f}(r^i) = \frac{1}{n}\sum_i \hat{f}(p|p_*^i). \tag{5}$$

3 Applications

The setting for the case study is a 2D channelized, highly heterogeneous reservoir. The geostatistical (fine grid) model is defined on a grid of size 120×120, covering an area of 1200×1200 ft^2. The permeability field is the only property that varies between different realizations of the reservoir variable. Other properties such as porosity, initial saturations and relative permeabilities are assumed known and are assigned the same values in every realization of the reservoir. For these simulations, the water-oil mobility ratio (based on endpoint mobilities) is 5. The use of a water-oil mobility ratio that differs from unity introduces nonlinearity into the flow equations.

3.1 GEOSTATISTICAL MODEL

One hundred realizations of the permeability field are drawn from the prior distribution for the reservoir variable, $f(r)$. Each realization is formed by first generating the two facies present, sand and shale. This is accomplished through multiple-point geostatistics by borrowing patterns from a training image using the SNESIM software (Strebelle, 2002). Each facies is then populated with log-Gaussian permeability fields with different means ($\mu_{sand} = 8$ md, $\mu_{shale} = 2.5$ md) and covariance structures ($\sigma_{sand} = 2.5$ md, $\sigma_{shale} = 1.25$ md), as shown in Figure 1. The only hard data used is information from the wells. Both the injection well and the production well are located in sand, and the actual values of the permeability in the well grid blocks are assumed known. The injection well is located in grid block (18,23) and the production well is located in grid block (103,103).

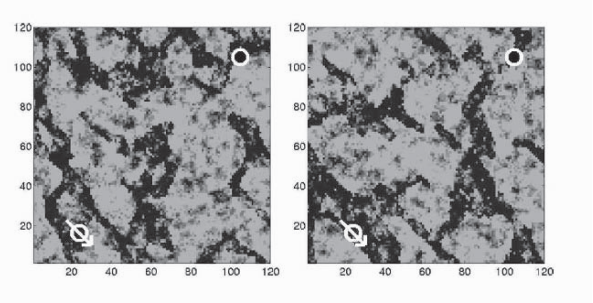


Figure 1. Two realizations of the permeability field. High permeability channels are displayed in black and low permeability shale in gray.

3.2 UPSCALING ALGORITHMS

The coarse scale simulations are performed on a coarsened mesh with properties computed using two different single-phase upscaling algorithms, namely purely local $\mathbf{k}^*$ upscaling and adaptive local-global T^* upscaling, as discussed below. Upscaling methods can be described in terms of the size of the region simulated in the determination of the coarse scale parameters (Chen et al., 2003). At one extreme are purely local methods, in which the fine scale region associated with a single target coarse block is simulated subject to a particular set of assumed boundary conditions. Extended local methods incorporate some amount of surrounding fine scale information into the calculations, though they still require assumptions regarding boundary conditions. Global techniques, by contrast, apply global simulations for the determination of upscaled quantities. A quasi-global procedure that uses specific coarse scale global simulations in conjunction with extended local calculations to determine the upscaled properties was recently introduced (Chen and Durlofsky, 2005) and will be applied here (this approach is referred to as adaptive local-global upscaling).

Upscaling algorithms may provide upscaled permeability (designated $\mathbf{k}^*$, where the * indicates an upscaled quantity), which is then used to compute the interblock transmissibilities required by simulators, or they may provide upscaled transmissibility (T^*) directly. In recent work, Chen et al. (2003) showed that direct T^* upscaling generally provides better accuracy than $\mathbf{k}^*$ upscaling for highly heterogeneous channelized systems. Upscaled near-well parameters (well index and well block transmissibilities) can also be computed and these can provide significant improvement in the accuracy of the coarse scale simulations in some cases (see e.g., Durlofsky et al., 2000).

The algorithms applied here are (i) a purely local $\mathbf{k}^*$ upscaling with standard pressure-no flow boundary conditions and no near-well upscaling and (ii) adaptive local-global T^* upscaling, which incorporates near-well upscaling directly in the algorithm. In both cases upscaling calculations are performed based on the single-phase (dimensionless) pressure equation:

$$\nabla \cdot (\mathbf{k}(\mathbf{x}) \cdot \nabla p) = q, \tag{6}$$

where $\mathbf{k}(\mathbf{x})$ is a spatially variable permeability tensor, p is pressure, and q is a source or sink term. From the discussion above, we expect better accuracy using adaptive local-global upscaling than with purely local upscaling. It is important to note, however, that purely local methods such as that applied here are widely used in practice. No two-phase upscaling (e.g., pseudo relative permeabilities) is applied in this work, so we anticipate some error in transport predictions when the grid is coarsened significantly.

In this study, the fine scale models are 2D and are reasonably small (120 ×120), so all of the fine scale simulations can be performed in a reasonable amount of time. The fine scale simulations that are not included in the calibration set are used to validate the procedure and results. This will not be possible in large 3D studies, so it is important that guidelines and procedures be established through idealized studies such as this.

3.3 CONSTANT TOTAL RATE CASE

In the first flow scenario, the injection well is kept at a constant injection rate, while the production well is kept at a constant bottom hole pressure (BHP). This results in a constant total rate (as the system is nearly incompressible). We present results in terms of the injection well BHP, which varies in time due to the contrast in mobility between oil and water. We first apply purely local $\mathbf{k}^*$ upscaling and coarsen the fine grid geostatistical model (120×120) to a grid of 24×24. One hundred realizations are considered and flow simulations are performed on all of the fine and coarse models.

Results for the 100 coarse simulations are shown in Figure 2 (left). The estimated mean and 90% confidence interval for the BHP in the injection well for the 100 coarse scale models are compared to the same estimates based on the 100 fine scale simulation runs (Figure 2, right). The estimate based on the fine scale runs can be considered as an estimate of the true distribution $f(p)$. It is immediately apparent that the upscaling has introduced a severe bias, in addition to much wider confidence intervals. Note that the lower limit of the 90% confidence interval estimated based on the coarse scale runs just includes the expected value estimated from the fine scale runs (it is worth noting that the 80% confidence interval from the coarse runs does not include the fine scale expected value). The increase in uncertainty may be due to the fact that purely local $\mathbf{k}^*$ upscaling captures (or fails to capture) the effects of large-scale permeability connectivity to varying degrees (depending on the realization), which leads to a high degree of variation in the accuracy of the coarse scale results.

We now apply the error modeling procedure described above to improve the coarse scale predictions. The size of the calibration set (n_{fine}) is chosen to be 5. This means that five fine scale fluid flow simulations are used in the error modeling procedure. We emphasize that the choice of which fine scale runs to use must be made in the absence of any fine scale simulation results. The choice of the particular simulations to use for this calibration can have a large impact on the results. To illustrate this, we first present results using a very 'unlucky' choice for the calibration runs (this set was found by *maximizing* the size of the confidence intervals). The 'corrected' results for mean and confidence intervals using this set of calibration runs is shown in Figure 3 (left). This result is actually worse than the initial (uncorrected) estimates.

We next apply the minimum variance criterion described above for the selection of the (n_{fine}) fine scale calibration runs. The corrected coarse scale results using this approach are shown in Figure 3 (right). Here we see that the corrected mean and upper confidence interval are very close to the fine scale results. There is still some error in the lower confidence interval, but the overall predictions show a very significant improvement relative to the uncorrected coarse scale results in Figure 2 (right).

The "unlucky" calibration set tends to include coarse scale values that are close together, resulting in a regression that is quite unstable. By using the minimum variance criterion, we select values that are spaced further apart, which provides a more stable regression. Since we determine the calibration set using more than one production variable (at many different time steps), we do not select only the most extreme values (as would likely occur if we based the regression on a single production variable).

We next demonstrate the potential impact of using a more accurate upscaling procedure. Shown in Figure 4 (left) are results using the adaptive local-global T^* upscaling approach. With this technique, the (uncorrected) estimates for the mean and confidence interval are quite close to those from the fine scale models, though a slight bias is evident. Using the minimum variance selection procedure with $n_{fine} = 5$ provides the corrected results shown in Figure 4 (center). Note the improved accuracy in the lower confidence interval relative to that achieved with purely local $\mathbf{k}^*$ upscaling (Figure 3, right).

An advantage of using a more accurate upscaling is that fewer fine scale runs can be used for the calibration. This represents significant potential computational savings, as the fine scale simulations will likely represent the largest computational component of our procedure in practical applications. Shown in Figure 4 (right) are results using adaptive local-global T^* upscaling for the coarse runs but with only 3 fine scale calibration runs. These results are still quite accurate and in fact display better accuracy than was achieved with purely local upscaling and $n_{fine} = 5$.

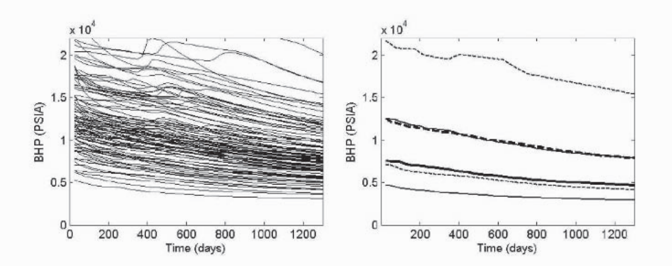


Figure 2. Left: BHP in the injection well for the 100 coarse scale models generated using purely local $\mathbf{k}^*$ upscaling. Right: Mean (thick dashed line) and 90% confidence interval (thin dashed lines) estimated based on these 100 coarse models compared to the mean (thick solid line) and 90% confidence interval (thin solid lines) estimated from the fine scale simulations.

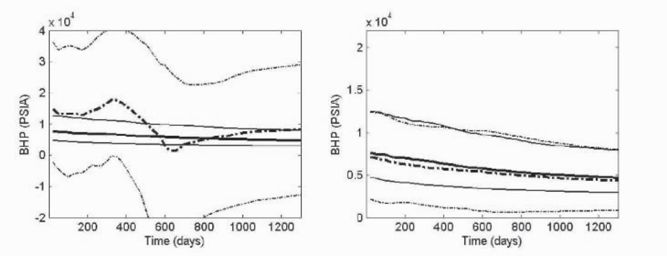


Figure 3. Left: Corrected mean (thick dot-dash line) and 90% confidence interval (thin dot-dash lines) for the 100 coarse models generated using purely local $\mathbf{k}^*$ upscaling with a particularly 'unlucky' calibration set. Right: Corrected mean and 90% confidence interval using the minimum variance calibration ($n_{fine} = 5$ in both cases). The mean (thick solid line) and 90% confidence interval (thin solid lines) estimated from the fine scale simulations are also shown in both figures.

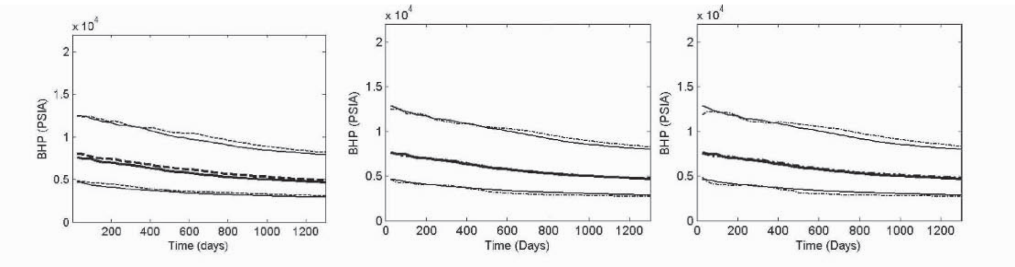


Figure 4. Left: Mean (thick dashed line) and 90% confidence interval (thin dashed lines) estimated based on the 100 coarse models generated using adaptive local-global T^* upscaling. Center: Corrected mean (thick dot-dash line) and 90% confidence interval (thin dot-dash lines) based on the minimum variance calibration set with $n_{fine} = 5$. Right: Same as center, but with $n_{fine} = 3$. The mean (thick solid line) and 90% confidence interval (thin solid lines) estimated from the fine scale runs are shown in all figures.

These results are also more accurate than those achieved with $n_{fine} = 7$ and purely local upscaling (not shown). This demonstrates the potential advantages of accurate upscaling within the context of our error modeling procedure. We note that the least number of realizations that can be used for the calibration is 3, since the linear regression in Eq. 4 requires 3 pairs of observations to determine a fit.

3.4 CASE WITH VARIABLE WELL CONTROL

In the second flow scenario, the injection well is kept at a constant BHP, while the production well is initially specified to produce at a constant oil rate. When the BHP in the production well drops to 1000 psi, the producer switches to BHP control. This case is more complex because the total rate must increase once water breaks through (to maintain the specified oil rate) and because the producer eventually switches to BHP control. It is often difficult to capture accurately discrete events such as these in upscaled models, so we expect to see higher levels of error in the uncorrected coarse scale simulations. In addition, the permeability fields are upscaled to a greater degree, from the initial 120×120 grid to a 12×12 grid. The fine scale realizations, well locations and system properties are the same as in the previous example. In this case we only present results for adaptive local-global T^* upscaling.

Results for the mean and confidence interval for the fine and uncorrected coarse models are shown in Figure 5 (left). The coarse scale predictions are accurate at early time, but lose accuracy as time progresses, particularly around water breakthrough (at approximately 300 days). The errors in this case are considerably larger than in the previous example, due to the switches in well control and also to the coarser grid. Using the minimum variance procedure to select fine scale calibration runs, we achieve the corrected results shown in Figure 5 (center) for $n_{fine} = 5$ and in Figure 5 (right) for $n_{fine} = 3$. The results are slightly more accurate with $n_{fine} = 5$ than with $n_{fine} = 3$, though the results are acceptable even with $n_{fine} = 3$. As was the case for the previous example, results with purely local $\mathbf{k}^*$ upscaling (not shown here) are considerably less accurate than those with adaptive local-global T^* upscaling. This example demonstrates the ability of the overall procedure to provide reliable results in a complex case involving a high degree of upscaling and variable well control.

4 Conclusions

This study demonstrated that upscaling procedures can introduce severe biases and other error structures into coarse scale flow simulation results. These errors can, however, be reliably modeled using procedures described here. This modeling requires that a few flow simulations be performed both at the fine and coarse scales. The realizations to be simulated at the fine scale must be selected carefully, and we described and illustrated a minimum variance criterion that provides an appropriate calibration set. By combining coarse scale simulations and the error model, the accuracy in the forecasts for both mean and confidence interval is greatly improved. It was also demonstrated that the use of a more advanced

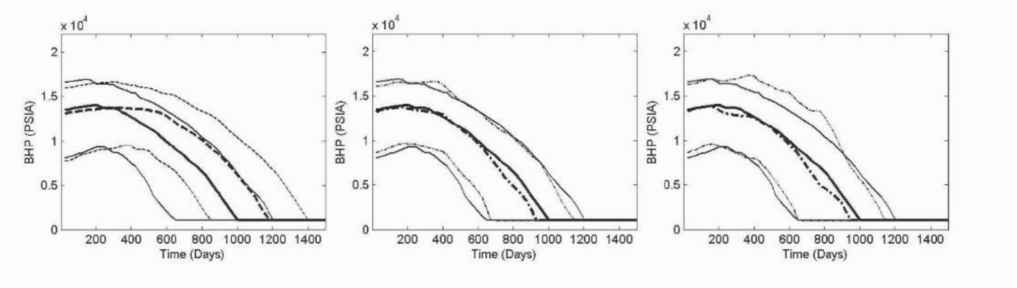


Figure 5. Case with variable well control. Left: Mean (thick dashed line) and 90% confidence interval (thin dashed lines) estimated based on the 100 coarse models generated using adaptive local-global T^* upscaling. Center: Corrected mean (thick dot-dash line) and 90% confidence interval (thin dot-dash lines) based on the minimum variance calibration set with $n_{fine} = 5$. Right: Same as center, but with $n_{fine} = 3$. The mean (thick solid line) and 90% confidence interval (thin solid lines) estimated from the fine scale runs are shown in all figures.

upscaling algorithm (adaptive local-global T^* upscaling) allows for the use of fewer fine scale simulations, which improves the computational efficiency of the overall procedure.

Acknowledgements

This study was funded in part by the Uncertainty in Reservoir Evaluation (URE) research program at the Norwegian University of Science and Technology and by the industrial affiliates of the Stanford University Reservoir Simulation Research Program (SUPRI-B).

References

Chen, Y. and Durlofsky, L.J., *Adaptive Local-Global Upscaling for General Flow Scenarios in Heterogeneous Formations*, to appear in Transport in Porous Media, 2005.

Chen, Y., Durlofsky, L.J., Gerritsen, M. and Wen, X.H., *A Coupled Local-Global Upscaling Approach for Simulating Flow in Highly Heterogeneous Formations*, Advances in Water Resources, vol. 26, 2003, p. 1041-1060.

Craig, P.S., Goldstein, M., Rougier, J.C. and Seheult, A.H., *Bayesian Forecasting for Complex Systems using Computer Simulators*, Journal of the American Statistical Association, vol. 96, no. 454, 2001, p. 717-729.

Durlofsky, L.J., Milliken, W.J. and Bernath, A., *Scaleup in the Near-Well Region*, SPE Journal, March 2000, p. 110-117.

Hegstad, B.K. and Omre, H., *Uncertainty in Production Forecasts Based on Well Observations, Seismic Data, and Production History*, SPE Journal, December 2001, p. 409-424.

Kennedy, M. and O'Hagan, A., *Bayesian Calibration of Computer Models*, Journal of the Royal Statistical Society, Series B, vol. 63, 2001, p. 425-464.

Omre, H. and Lødøen, O.P., *Improved Production Forecasts and History Matching using Approximate Fluid Flow Simulators*, SPE Journal, September 2004, p. 339-351.

Strebelle, S., *Conditional Simulation of Complex Geological Structures using Multiple-point Statistics*, Mathematical Geology, vol. 34, no. 1, 2002, p. 1-22.

SENSITIVITY OF OIL PRODUCTION TO PETROPHYSICAL HETEROGENEITIES

ARNE SKORSTAD[1], ODD KOLBJØRNSEN[1], BJØRN FJELLVOLL[1], JOHN HOWELL[2], TOM MANZOCCHI[3] and JONATHAN N. CARTER[4]
[1]*Norwegian Computing Center, PO Box 114 Blindern, NO-0314 Oslo, Norway.* [2]*University of Bergen, Norway.* [3]*Fault Analysis Group, University College Dublin, Ireland.* [4]*Imperial College, London, UK*

Abstract. The multi-partner Saigup project was established to systematically investigate the relative importance of key geological heterogeneities on simulated production. In order to investigate the impact of the geostatistical variability, the petrophysical fields were drawn repeatably, and the variogram directions were rotated. The heterogeneities caused by the geostatistical variability in the petrophysical simulations and the variogram rotation were similar, and had a low impact on all the production responses except total water injected. Here they contributed about 20% of the total variability.

1 Introduction

Production from a reservoir is a complex function of many parameters. Reservoir modelling enables the prediction of reservoir performance through mathematical simulations of the flow. However, these deterministic flow simulations include significant uncertainties, due mainly to uncertainty in the original geological input in the reservoir models. Geostatistical models should aim to deal with this uncertainty.

The objective of the European Union supported Saigup project was to quantify the effects of geological variability and the associated uncertainty in faulted, prograding shallow marine reservoirs. Related, smaller scale studies are e.g. Lia et al. (1997) and Floris et al. (2001). Within Saigup, a series of geological parameters were varied systematically in order to reveal their relative importance on reservoir production. By comparing different realizations of the petrophysical fields, the stochastic variability is also quantified, c.f. Manceau et al. (2001). For a subset of the realizations the lateral petrophysical variogram anisotropy direction was rotated. This set gave information on how significant variogram direction is for the total variability in simulated production data.

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 723–730.

2 Saigup variables

The synthetic Saigup reservoirs are a 3 x 9 km prograding shallow marine tilted fault block, comprised of four 20 m thick parasequences (Figure 1). Each parasequence contains up to six facies associations ranging from offshore through delta-front to coastal plain with channels. The facies were populated with petrophysical properties drawn from distributions taken from comparable North Sea (mainly Brent Group) reservoirs.

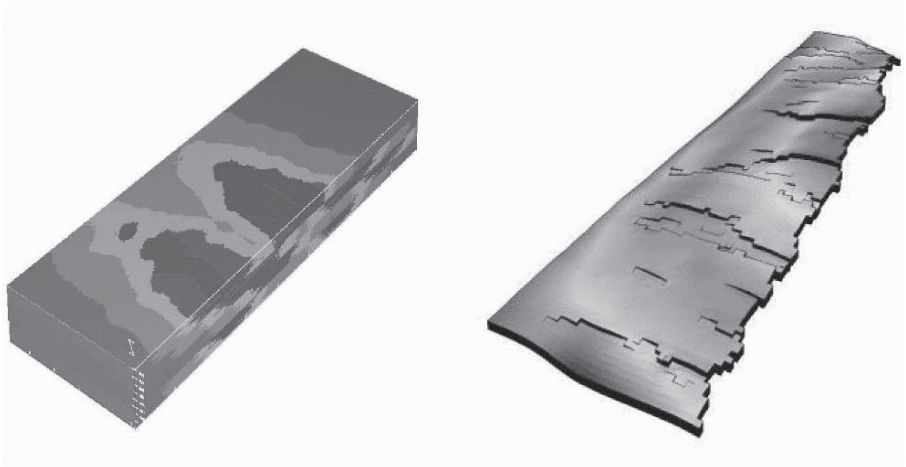


Figure 1. Facies representation of a synthetic Saigup reservoir with eastward progradation direction (left). The reservoir is then aligned under the top structure map (example with strike perpendicular fault pattern to the right). North is to the upper right edge.

Production heterogeneity was introduced by varying eight different parameters at three different levels. Seven of these are related to geology and one to production strategy. These variables included four sedimentological parameters: aggradation angle (shoreface trajectory); progradation angle relative to waterflood; delta type reflected in shoreline curvature and internal flow-barrier coverage. Structural parameters were fault permeability, fault pattern (unfaulted, compartmentalized, strike parallel and perpendicular to main flow) and fault density.

The final parameter that was varied was the well pattern. Four different well configurations (designed for each structural pattern) were run on all of the geological models. The well configurations were a combination of vertical producers and injectors. The producers were located at a high level near the crest, see Figure 2, while injectors were located at a lower level, injecting water subject to a maximum pressure of 50 bar above the initial reservoir pressure.

The first stage of the workflow was a geostatistical simulation of the sedimentology based on the chosen control levels. This produced a facies model that guided the subsequent geostatistical simulation of the petrophysical parameters.

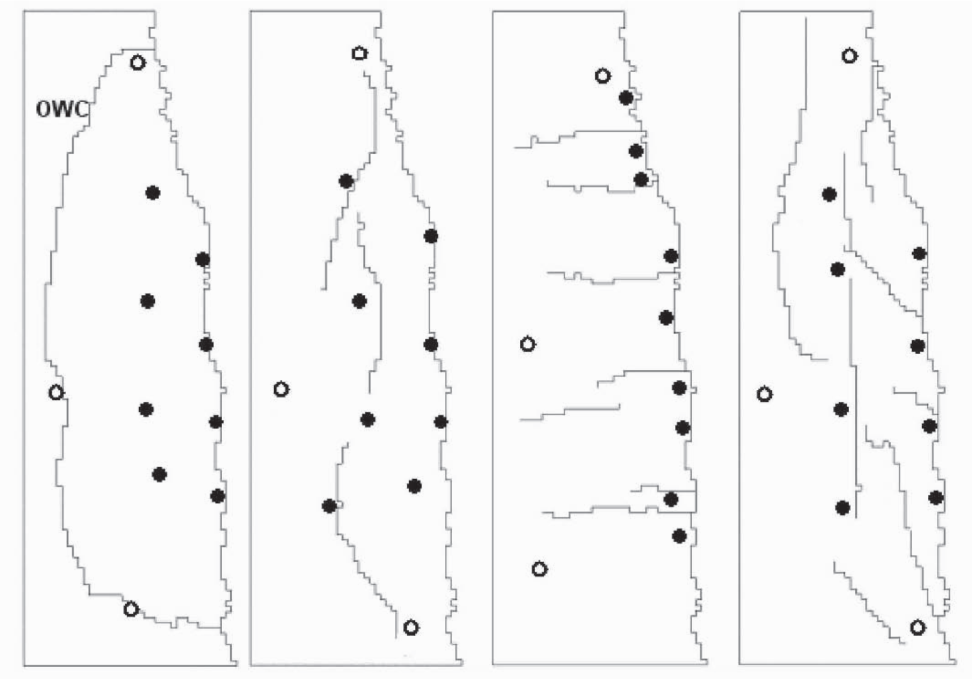


Figure 2. View of well locations designed for, from left to right, unfaulted, strike parallel, strike perpendicular and compartmentalized fault patterns. Injectors are shown as circles and producers as dots. North is to the upper edge. Only the largest faults are shown, and the owc is indicated in the unfaulted case.

The fine scale petrophysical model was then upscaled from 1.5 million geogrid cells to 96 000 flow-simulation cells. The upscaling method was a local flow preserving method with open boundaries, commonly used by the petroleum companies (Warren & Price, 1961). Relative permeability upscaling sensitivity was investigated by Stephen et al. (2003). The flow-barriers were simulated as transmissibility multipliers within the simulation grid. Finally the fault related heterogeneities were introduced and the realization was ready for flow-simulation. The flow-simulator was run for 30 years (production time) producing the final production responses.

All combinations of the 8 input key controls were run with repeated simulations on some of the combinations. In total, more than 12 000 flow simulations relevant for this analysis were run.

3 Variance components

A production response variable is a function of its explanatory variables. In the Saigup study there were originally eight key control parameters: four sedimentological, three structural and one well related explanatory variable. Also two geostatistical variables were investigated: the variogram anisotropy direction related (V) and the repeated petrophysical simulation effect (P). For simplicity, all the sedimentological, structural and well related variables are merged here into one geoscience variable G. The production response y can then be written

$$y(G, P, V) = K_0 + K(G, P, V), \tag{1}$$

where the average level is K_0 and the function $K()$ describes the variation around that average. In order to investigate the different explanatory effects by statistical analysis, this is broken down to its orthogonal effects

$$\begin{aligned} K(G, P, V) \;=\; & K_G(G) + K_P(P) + K_V(V) + K_{G,P}(G, P) + \\ & K_{G,V}(G, V) + K_{P,V}(P, V) + K_{G,P,V}(G, P, V). \end{aligned} \tag{2}$$

Thereby it is possible to quantify the relationship between the variability of the different explanatory variables and the response by separating the variance components. Estimates are obtained by a standard moment method, cf. Box et al. (1978).

4 Effect from repeated stochastic simulation of petrophysics

We are interested in the relative contribution of the geostatistical variability from the petrophysics (P) in equation (2). That is the variability obtained by changing the seed in the petrophysical simulation. The relative effect of the repeated stochastic simulation depends on the main effect and all higher order combined effects,

$$E_P^r = \sqrt{\frac{||K_P||^2 + ||K_{G,P}||^2 + ||K_{P,V}||^2 + ||K_{G,P,V}||^2}{||K||^2}}. \tag{3}$$

The results for the selected responses are given in Table 1. With the exception of the total water injected response, the effect is very low. This is because $||K_G||$ dominates equation (2). The reason why the water injection is more subjected to changes in the petrophysical field than the oil production, is believed to originate from the different flow characteristics of the two fluids.

Table 1. Relative effect E_P^r of repeated petrophysical fields on variability in production responses.

Production response	Relative effect
Total oil production	1.1%
Discounted production	1.3%
Recovery factor	3.8%
Recovery at 20% pore volume injected	3.3%
Total water injected	15.8%

5 Variogram direction effect

Reservoir properties produced by the deposition of sediment within a prograding shallow marine system will not be horizontally isotropic. Heterogeneity within the distributary channel sediments will be aligned broadly normal to the channel belt orientation. Within shallow marine deposits the greatest heterogeneity will occur perpendicular to the shoreline orientation (Kjønsvik et al., 1994); (Miall and Tyler, 1991). In the Saigup study, the variograms of the six different facies were all spherical. For the two most permeable facies associations, the variogram ranges were 800 and 250 meters for the channels, and 2000 and 1000 meters for the upper shoreface facies.

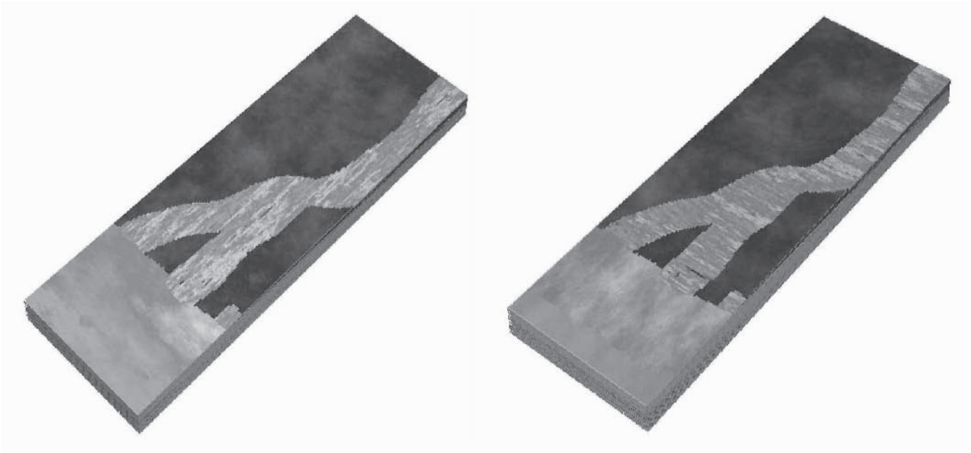


Figure 3. Rotation of variogram direction of petrophysical parameter for a southward prograding realization.

In order to investigate the importance of correct variogram anisotropy direction, a new series of repeated petrophysical simulations were made using identical facies realizations, now with 90° rotated petrophysical variograms, see Figure 3. The additional effect from rotating the variogram anisotropy direction is

$$\Delta E_V^r = \sqrt{\frac{||K_V||^2 + ||K_{G,V}||^2}{||K||^2}}, \tag{4}$$

where equations (2) and (3) give that $(E_P^r)^2 + (\Delta E_V^r)^2 = 1 - ||K_G||^2/||K||^2$. Consequently there were no difference with respect to the element K_G, and any differences in the variance components are therefore due to the other elements of equation (2). These estimates are shown in Table 2. The values are low which indicates that the variogram direction has little impact. Note also that the higher order contributions from $K_{P,V}$ and $K_{G,P,V}$ are included in equation (3). A natural conclusion is that both of the investigated geostatistical variabilities (Table 1 and Table 2) are relatively small.

Table 2. Relative effect ΔE^r_V of petrophysical variogram direction rotation on variability in production responses.

Production response	Relative effect
Total oil production	0.2%
Discounted production	0.0%
Recovery factor	0.0%
Recovery at 20% pore volume injected	0.5%
Total water injected	5.3%

Unlike other applications (such as mining) where correct anisotropy direction are important, the petroleum industry does not deal with single properties of the rock itself, but on a complex fluid flow function which is controlled by a variety of rock properties acting at different scales. Because of this, apparent errors in the orientation of the variogram direction may not be as crucial. In fact by doing so, the fingering is reduced in the model, the amount of produced water is reduced, and the sweep efficiency is increased. So it may take longer time for the fluids to get to the producer in the simulated model, but more valuable fluids will reach the producer before the water-cut becomes too high. The results suggest however that this effect is low compared to the effects of other uncertain input parameters.

6 Discussion

The effect of the stochastic variability on the computed oil production rate is illustrated in Figure 4. The rates are in accordance with the low effect seen in Table 1 and Table 2.The differences in the production curves within each reservoir are much smaller than the main features of the productions. The means and standard deviations of the cumulative production were 388 and 18 MSM3 (million standard cubic meters) for the upper and 277 and 7.5 MSM3 for the lower reservoir respectively.

The upper reservoir has an early high production rate which becomes much lower after 15 years, while the lower reservoir remains on a lower plateau for much longer. These differences are due to uncertainties in the geological model which determines most of the variability in the production response of the Saigup reservoirs. The variability from repeated geostatistical petrophysical simulations is comparable to that for the rotated variograms. This was also observed in other synthetic Saigup reservoirs. The importance of the stochastic variabilities will however be more significant if the uncertainties in the key geological parameters are reduced.

The observation that the effects of the two geostatistical variabilities considered here are quite small compared to those produced by geological variability is important. It illustrates that efforts should be focused on dealing with uncertainties in geological parameters that are key to describing the reservoirs.

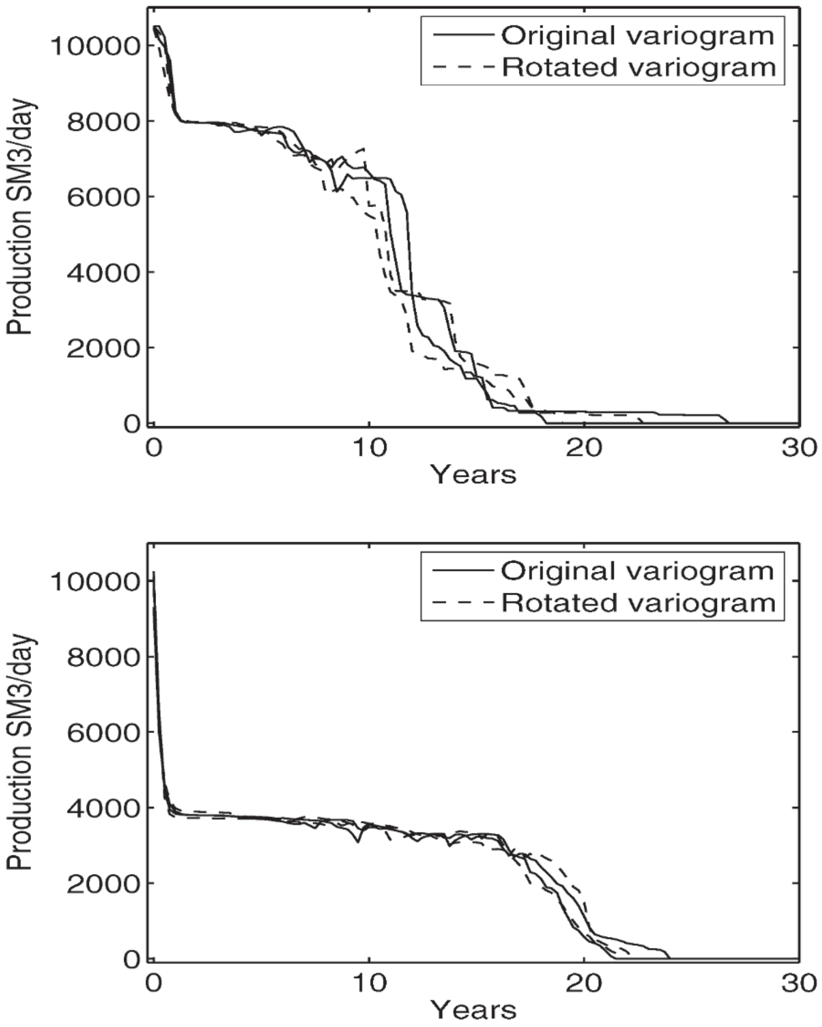


Figure 4. Four production rates observed from two reservoirs. Each line type shows the effect of repeated petrophysical stochastic simulations. The dashed curves have rotated variogram anisotropy directions compared to the solid curves.

The simulated production data indicates that if the ratio between the variogram anisotropy ranges is below 3, the actual anisotropy directions are not crucial for the cumulative response in prograding shallow marine reservoirs. Other uncertainties are far more dominant, and this variability is comparable to that of the geostatistical uncertainty originating from the repeated stochastic simulation (changing the seed).

The applicability of these results to real world reservoirs is dependent upon how representative Saigup parameter space is of the real world. Significant care was taken to ensure that the initial collection of data covered realistic geological parameter ranges and consequently we believe this has been addressed.

Acknowledgements

The Saigup project has been a truly integrated study, with contributions by colleagues from all the ten project partners. The authors therefore wish to acknowledge that contribution from University College Dublin (IRL), Imperial College (UK), Heriot-Watt University (UK), University of Liverpool (UK), Norwegian Computing Center (NO), TNO-NITG (NL), Shell (SIEP-SEPTAR) (NL), BG (UK), Badley Earth Sciences (UK) and Roxar Software Solutions (NO). We also thank the European Union for financially supporting this project.

References

Box, G. E. P., Hunter, W. G. and Hunter, J. S. *Statistics for experimenters.* John Wiley & Sons, New York. 1978.

Floris, F. J. T., Bush, M. D., Cuypers, M., Roggero, F. and Syversveen, A-R. *Methods for quantifying the uncertainty of production forecasts: a comparative study.* Petroleum Geoscience, vol. 7. Special issue on geostatistics, S75-S86. 2001.

Kjønsvik, D., Doyle, J. and Jacobsen T. *The effects of sedimentary heterogeneities on production from shallow marine reservoirs – what really matters?* European Petroleum Conference, London, 25-27 October 1994. SPE 28445.

Kolbjørnsen, O., Skorstad, A., Holden, L., Howell, J., Manzocchi, T. and Carter, J. N. *Influence on Geological Factors on Oil Production in Shallow Marine Reservoirs.* 1st EAGE North Africa/Mediterranean Petroleum & Geoscience Conference & Exhibition. S051. Tunis, Tunisia, 6 - 9 October, 2003.

Lia, O., Omre, H., Tjelmeland, H., Holden, L. and Egeland, T. *Uncertainties in reservoir production forecasts.* AAPG Bulletin, Vol. 81, Nr. 5, 1997.

Manceau, E., Mezghani, M., Zabala-Mezghani I. and Roggero, F. *Combination of Experimental Design and Joint Modeling Methods for Quantifying the Risk Associated With Deterministic and Stochastic Uncertainties – An Integrated Test Study.* Proc. 2001 SPE Annual Technical Conference and Exhibition, New Orleans, 30 September - 3 October. SPE 71620.

Miall, A. D. and Tyler N., eds. *The three dimensional facies architecture of terrigenous clastic sediments and its implications for hydrocarbon discovery and recovery.* Soc. Econ. Paleont. Mineral.

Stephen, K. D., Yang, C., Carter, J. N., Matthews, J. D. and Howell, J., *Sensitivity Study of Facies Based Geopseudo Two-Phase Upscaling in a Shallow Marine Environment.* 65th EAGE Annual Conference & Exhibition. P038. Stavanger, Norway, 2 - 5 June, 2003.

Warren, J. J., and Price, H. S., *Flow in Heterogeneous Porous Media.* SPE Journal, September, 1961.

SCALING RELATIONS AND SAMPLING VOLUME FOR SEISMIC DATA

PETER FRYKMAN, OLE V. VEJBÆK and RASMUS RASMUSSEN
Geological Survey of Denmark and Greenland - GEUS
Øster Voldgade 10, DK-1350 Copenhagen K, Denmark

Abstract. Seismic attributes are considered valuable auxiliary data for geostatistical reservoir modelling. The seismic resolution is often poorly defined and depends on acquisition, depth and processing scheme. In this study an approach to determine the practical sampling volume of the seismic data is described, exploiting the geostatistical scaling laws.

The analysis sequence involves calculating the synthetic seismic response from a fine-scale model originating from a chalk reservoir; inversion of the generated data into impedances, and finally a quantitative comparison of this derived synthetic seismic attribute with the original fine-scale data.

The sampling volume represented by the inversion data is then estimated from the comparison of the volume-variance and the variogram structure. The derived sampling volume is expressing a generic resolution, since added noise and variance from real-life acquisition and processing is absent, and the result therefore only provides a bounding limit for the practical resolution.

1 Introduction

The recent advances in reservoir modelling have increased the demand for quantitative description and higher resolution in the geological model, and thereby highlight the scaling issues, although most papers ignore the question. The available data almost always measure a different volume scale than the volume of the grid cells used in the numerical model; therefore a strategy to reconcile these differences must be developed. For modelling of facies, porosity, or other reservoir properties, different data types are used for a variety of cosimulation schemes. Therefore core data, well log data, seismic data, and even well test data often contribute to the same reservoir model. It is increasingly recognised that the volume-variance relations as described in the scaling laws link these different data types, and must be considered prior to incorporation in the modelling procedure.

It has become nearly standard in reservoir modelling to use data derived from seismic inversion results or seismic attribute analysis. As these measures are often combined with other data at either finer scale (core, well logs) or coarser scale (e.g. well tests), it is important to know more precisely what length/volume scale is represented. If the grid cell size of the reservoir model is different from the volume represented by the data, this must be accounted for.

Reservoir modelling in a geostatistical scheme requires the input variogram and the target histogram to be representative for the scale of the modelling cells. Therefore one must be able to estimate these two properties for the different data types, that might be

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 731–736.

derived at different scales. It has been shown that for simple measures such as porosity, the geostatistical scaling laws can be used to transcend different scales, and can calculate the variogram and histogram at any desired scale (Frykman & Deutsch 2002).

2 RESOLUTION OF SEISMIC DATA

An illustration of the different scales shows that the change of scale from core to log measurement volumes is as large as the jump from log volume to that of a geological modelling cell (Figure 1). The seismic detection volume is generally considered to lie within the range of the different modelling cell sizes, but it should be remembered that there is a large difference between the vertical and the horizontal resolution of the seismic data.

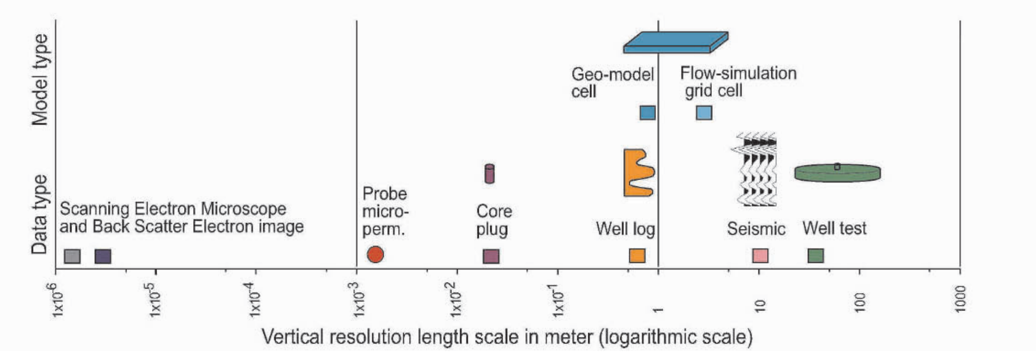


Figure 1. Illustration of the vertical length resolution measures (in meters) for the different types of data and model types.

2.1 The traditional estimate of seismic resolution

When a first rough estimate of the vertical seismic resolution is considered, it is necessary to assume an average sound velocity and a dominant frequency content in the particular section that is under investigation. For a specific case like a North Sea chalk reservoir at 2 km depth, the sound velocity of the chalk will be approximately 3000 m/sec and the dominant frequency around 60 Hz. This results in a wavelength of 50 m, and the best possible vertical resolution of features is therefore normally assumed to be around ¼ to 1/8 of this, i.e. 12.5 to 6 meter at best. The 6 meter is therefore often chosen as the vertical sample size in studies of these reservoirs, and corresponds approximately to the 4 msec TWT sampling interval used in the traditional seismic analysis.

3 THE GEOSTATISTICAL SCALING LAWS

The scaling laws indicate how statistics such as the histogram and variogram change with the volumetric scale (Kupfersberger *et al.* 1998; Frykman & Deutsch 2002). As scale increases, the range of correlation increases, the variance and variogram sill decrease, and the nugget effect also decreases. The main principles and examples with application of the scaling laws to porosity data have been described (Frykman & Deutsch 2002).

3.1 Brief recall of the volume-variance scaling laws

Denoting a smaller volume by $|v|$ and a larger volume by $|V|$, the two most important definitions are outlined:

A) The variogram *range a* increases as the size of the sampling volume increases, and a comparison of different scales is therefore dependent on the difference between the volumes, or |*V*| - |*v*|. Note that |*V*| and |*v*| relates to a size of the volume in a particular direction. Then, if a_v is the range for the small scale and a_V is the range for the larger scale, we have:

$$a_V = a_v + (|V| - |v|) \qquad (1)$$

B) The variance contribution, or *sill*, of each structure C_V^i, change by

$$C_V^i = C_v^i \frac{1-\overline{\Gamma}(V,V)}{1-\overline{\Gamma}(v,v)} \qquad (2)$$

$\overline{\Gamma}(v,v)$, or the *gamma-bar* value, represents the average variogram for vectors where each end of the vector independently describes the volume *v*. In 3D the gamma-bar values may be expressed by the infamous sextuple integrals of early geostatistics (Journel and Huijbregts 1978, p. 99). The modern approach, however, is to calculate all gamma-bar values numerically.

The variogram used for the calculation of the average variogram values, $\overline{\Gamma}$, is the unit point scale variogram Γ (i.e. variogram with sill of 1.0 and point range a_p).

The scaling relations are established under the assumptions that: 1) the shape of the variogram (i.e., spherical, Gaussian, etc.) does not change; 2) the averaging is performed with non-overlapping volumes, and 3) the variable scales in a linear fashion (Journel & Huijbregts 1978). The last assumption has so far prevented the use of the conventional scaling laws on parameters that scale non-linearly, like e.g. permeability, and the two other assumptions are questionable for parameters like acoustic impedance.

Obviously, these assumptions must be relaxed somewhat in order to be applied on our present case of seismic impedances, since a rigorous scaling should really be performed in the frequency/time domain. Our relaxation in this specific chalk case study is based on the fact that the contrasts in the chalk are limited, and therefore the effects of different averages are considered minor. The physical averaging by the seismic wave is performed in the continuos time domain, and cannot be claimed to reflect non-overlapping volume averaging. The importance of this assumption is unknown, but could have the effect of changing the variogram shape from spherical to Gaussian as it was shown for a simple moving window averaging (Frykman & Deutsch 2002).

4 CASE STUDY WITH SYNTHETIC SEISMIC DATA

An interval of 250 meter (2150 - 2400 m TVD) was selected from a well in the Upper Cretaceous chalk section in the Danish North Sea. Figure 2 shows the log data for both the porosity and the impedance which seem to be fairly stable without any large overall trends. As we are interested in the correlation structure at meter and 10-meter scale, this section length is deemed sufficient for our analysis.

The impedance data at well log scale is used to generate a synthetic seismic trace by applying a standard Ricker wavelet. The synthetic seismic trace is then inverted with standard procedure, including the use of a low-frequency model to keep the major trends in the inverted impedance trace (Figure 2). All impedance data used in this example are treated in the units of 10^6*Kg*m^{-2}*s^{-1}.

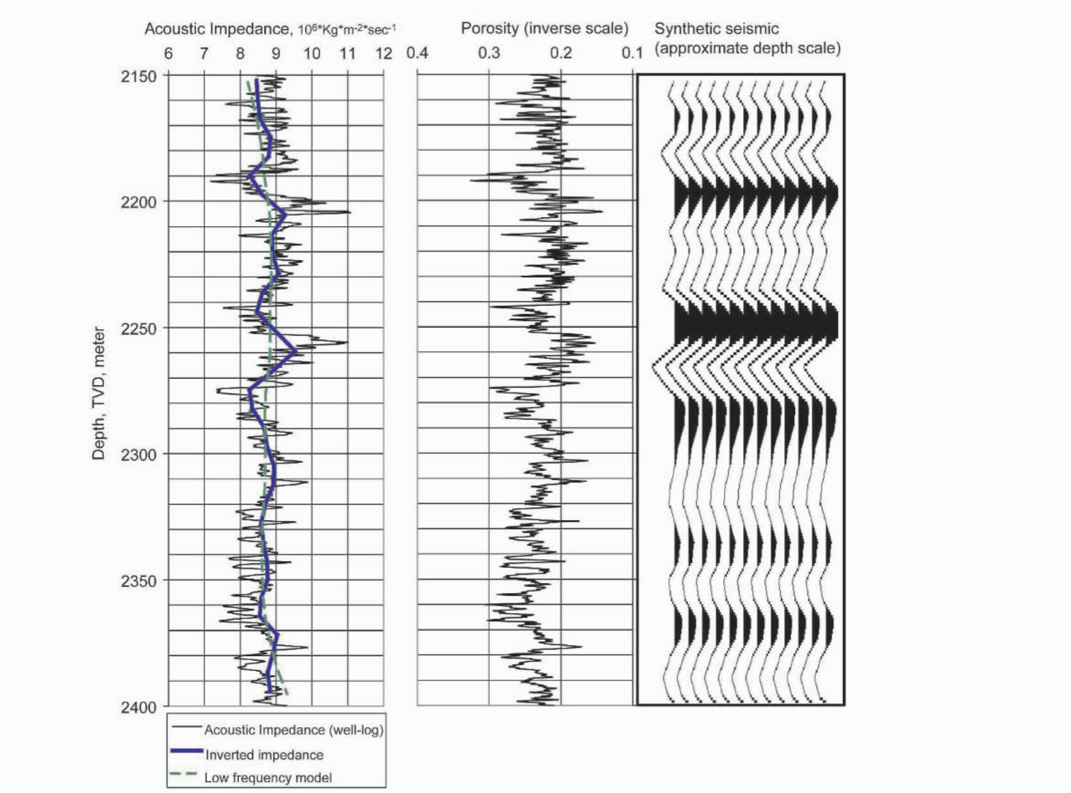


Figure 2. Impedance and porosity data from the well in a chalk reservoir. The fine-scale impedance is used to generate the synthetic seismic trace, which is inverted back to impedance including the use of a low-frequency model.

4.1 Analysis of the inverted impedance data

Obviously, the inverted impedance trace has a much lower variance than the original well log data. The difference in variance is the main entrance to derive the seismic resolution for the present case. The differences in correlation structure, i.e. the ranges and contributions from different structures, are adding to this analysis.

The variance for the inverted impedance is reduced to 0.071 compared to the 0.300 of the original log data (Figure 3).

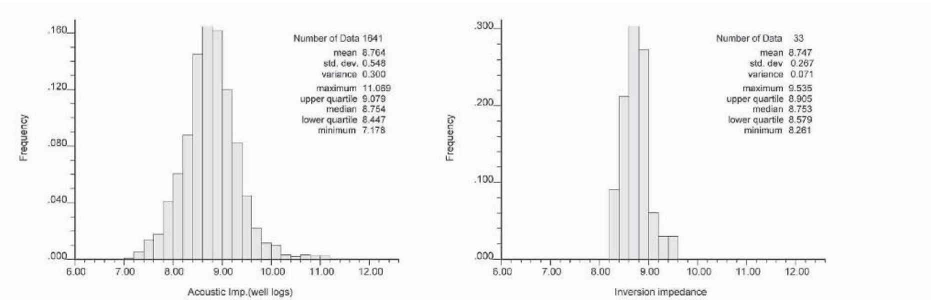


Figure 3. Comparison of histograms for impedance data.

By calculating the variance change for different volume sizes with the scaling laws using the average variogram method, it can be deduced for which upscaling length scale the variance in the inverted impedance is matching the theoretical variance (Figure 4). The log scale variogram used has two nested spherical structures with 1.6 and 11.0 m ranges, and nearly equal variance contribution.

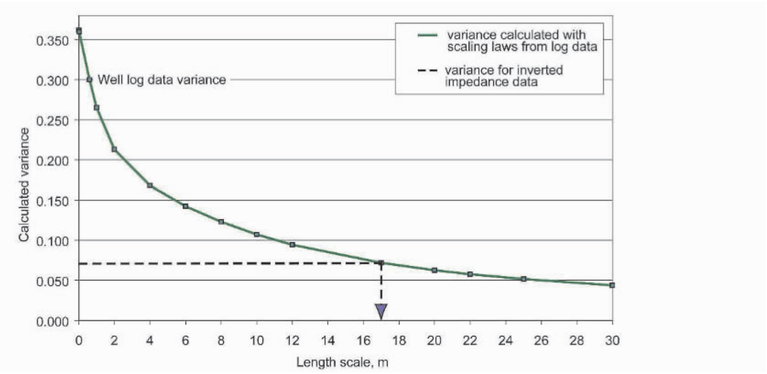


Figure 4. Variance as a function of the upscaling length, using the variogram structure for the original well log derived impedance data as a basis. It is seen that the experimental variance of the inverted impedance data of 0.071 corresponds to a length scale of 17.0 m.

For the present example, 17 m seems to describe the equivalent vertical upscaling volume for the seismic data with respect to the variance present in the inversion data. When this length is then used for scaling of the variogram structure from well log scale to 17 m seismic scale by using the scaling laws, the variogram model changes accordingly. The resulting upscaled variogram is not similar to the experimental variogram for the inverted impedance data (Figure 5A). When an additional cyclic component with a wavelength of 27 meter is included in the finer-scale model, the upscaled model is coming closer to the experimental variogram. Still, there is a mismatch between the ranges for the variograms.

If the starting point for the scaling investigation is taken from the inverted data, a single spherical variogram model with a range around 14 meter can be fitted to the inverted impedance variogram (Figure 5B). Using the scaling law for the range modification during downscaling, and assuming a log scale variogram with a single structure with 9.0 m range, an upscaling length measure of 5.6 meter is then deduced from the two variogram correlation lengths. A significant mismatch occurs in the variance if the 5.6 m is used as the seismic volume measure (Figure 5B).

5 Discussion

The discrepancy between the two values for the vertical seismic resolution derived from variance analysis and from variogram range analysis points to caution when using the scaling laws on inverted seismic data for downscaling to finer scales. The sequence of generated seismic (synthetic) and inversion illustrates in this case that the variance is reduced significantly more than predicted by the scaling laws, whereas the frequency content seems to match the predictions.

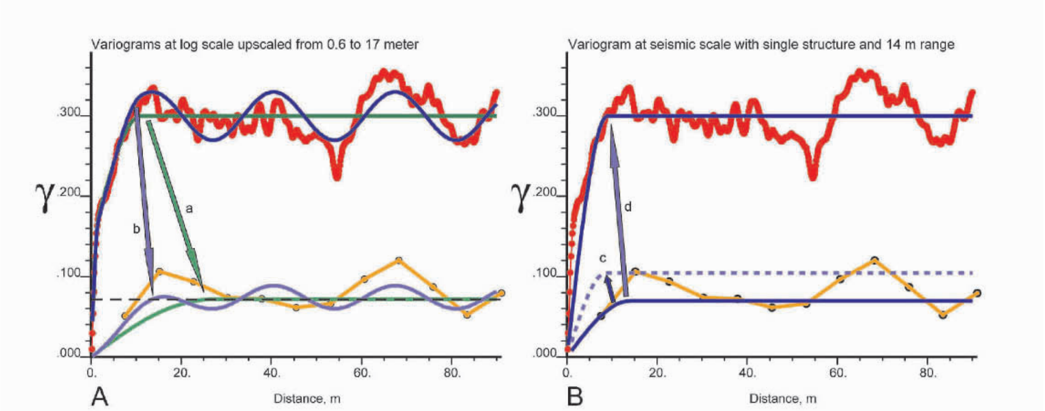


Figure 5. A: Variogram for the original impedance data at well log scale (red), with 2 alternative variogram models (green&blue). The simplest model (green) is a nested model with 2 spherical structures, and another model has an additional cyclic component as a hole effect structure (blue). The experimental variogram for the inverted impedance data (orange) and the two upscaled models (a, b, lightgreen and lightblue) are shown in the lower part of A.

B: Model variogram for seismic scale (blue) is downscaled (c) with scaling laws from 6 m to 0.6 m scale, and (d) variance modified to log-scale variance retaining the range.

6 Conclusions

The study illustrates how the vertical sampling volume represented by the inversion data can be estimated from the comparison of the volume-variance and of the variogram structure. The sampling volume of 17 meter in the vertical direction derived from variance analysis highlights the inability of the scaling laws to account for volume-variance relations in the frequency/time domain. The scaling of correlation range derives a value that matches traditional resolution estimates. The sampling interval of 4 msec TWT (approximately 7-8 m) normally used in the seismic analysis is therefore supported. The estimated resolution of 5.6 meter is a generic resolution, since added noise and variance from real-life acquisition and processing is absent. The result shown here provides a bounding limit for the practical resolution, which could be a larger length scale in other real cases.

Acknowledgements

The Geological Survey of Denmark and Greenland (GEUS) is thanked for permission to publish this paper.

References

Frykman, P. & Deutsch, C.V. Practical application of the geostatistical scaling laws for data integration. Petrophysics 43(3), 2002,153-171.

Journel, A.G. & Huijbregts, C. Mining geostatistics, New York City: Academic Press, 1978.

Kupfersberger, H., Deutsch, C.V. & Journel, A.G. Deriving constraints on small-scale variograms due to variograms of large-scale data. Mathematical Geology 30(7), 1998, 837-852.

HIDDEN MARKOV CHAINS FOR IDENTIFYING GEOLOGIC FEATURES FROM SEISMIC DATA

JO EIDSVIK
Statoil Research Center, 7005 Trondheim, Norway

EZEQUIEL GONZALEZ AND TAPAN MUKERJI
Department of Geophysics, Stanford Rock Physics Laboratory, Stanford University, California 94305

Abstract. In this paper we propose a hierarchical model for accomodating geologic prior knowledge together with velocity and density observations. The goal is to characterize underlying geologic patterns in clastic depositional systems, patterns such as blocky sand, shales, and fining or coarsening upward sequences. We use Gibbs sampling to explore the statistical distributions. The method is tested on synthetic data and well data from a fluvial environment.

1 Introduction

It is important that quantitative methods for lithology estimation from geophysical data honor geologic knowledge. However, it has been difficult to reconcile the mathematically formulated geophysical relations with qualitative geologic descriptions of lithologic sequences, such as blocky, fining or coarsening upwards, etc. One possible approach, that we focus on here, is to assume that the geologic variables of interest are governed by Markov transitions, see e.g. Weissman and Fogg (1999), Fearnhead and Clifford (2003), and Eidsvik, Mukerji, and Switzer (2004).

We present a hierarchical model for estimating underlying geologic patterns from P-wave velocity and density measurements in a well log. Since changes in velocity and density control seismic reflections, this preliminary study is a step towards assessing lithologic alternation styles from seismic data away from wells.

2 Methods

Our methods for inferring the underlying geologic environments and their alternation styles are illustrated in Figure 1. This hierarchical structure accommodates the relationships between hidden variables and the well measurements of velocity v_P and density ρ. The goal is to estimate the underlying geologic formations along the well path. The geologic classes are categorized as: 1=blocky sands, 2=fining upwards, 3=coarsening upwards, and 4=blocky shales.

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 737–742.

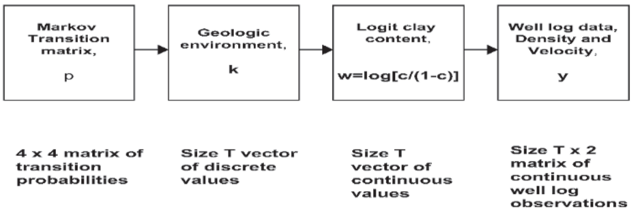

Markov Transition matrix, p
Geologic environment, k
Logit clay content, w=log[c/(1-c)]
Well log data, Density and Velocity, y
4 x 4 matrix of transition probabilities
Size T vector of discrete values
Size T vector of continuous values
Size T x 2 matrix of continuous well log observations

Figure 1. Display of a hierarchical model for the geophysical data. The direction of arrows indicates the conditional dependency. Variability is propagated from the left parameters to the observations at the right.

2.1 NONLINEAR MODEL FOR DENSITY AND VELOCITY IN SHALY SANDS

The observed density and P-wave velocity are measured at a constant interval for a total of T locations, and are denoted by $y_t = (v_{P,t}, \rho_t)$, $t = 1, \dots, T$. We assume that the lithology is a sand-shale mix described by the shaly-sand model of Marion et al (1992), in which density and velocity are nonlinearly connected to the underlying clay content fraction denoted by c_t. In this model ρ and v_P of the mixed shaly-sand are given in terms of the properties of the end members: pure sand and pure shale. The constants and parameters in the nonlinear calculations depend on the mineral densities (quartz and clay) and the end member porosities. Water saturation effect is taken into account using Gassmann's relation, see e.g. Mavko, Mukerji, and Dvorkin (1998). The relations are split into two domains: sand grain supported when clay content is less than the pure sand porosity ϕ_0, and clay matrix supported when clay content is greater than ϕ_0. The two domains correspond to the two limbs of a V-shaped response for ρ and v_P as a function of the clay content. Marion et al (1992) justify this split using physical interpretations along with lab measurements. The V-shaped trend has also been observed in log data. Other, similar functional relationships are presented in Koltermann and Gorelick (1995).

In a statistical formulation we define the probability density function (pdf) [conditional on logit clay content $w_t = \log(\frac{c_t}{1-c_t})$] for the data as:

$$f(y_t|w_t) = N\left[y_t; g\left(\frac{\exp(w_t)}{1+\exp(w_t)}\right), S\right], t = 1, \dots, T,\ f(y|w) = \prod_{t=1}^{T} f(y_t|w_t), \quad (1)$$

where $N(y; \mu, S)$ denotes the Gaussian pdf with mean μ and covariance S, evaluated at y, and where the two dimensional expectation term $g(c_t)$ is the nonlinear functional relationship in the shale-sand mix model of Marion et al (1992).

2.2 CLAY CONTENT SEQUENCE

We choose to use the logit clay content defined by $w_t = \log(\frac{c_t}{1-c_t}) \in \mathcal{R}$, $t = 1, \ldots, T$ in our modeling. The conditional pdf is denoted $f(w|k)$, where we assign a Markov property to w_t, conditional on the underlying geologic variables k_t;

$$f(w_t|w_{t-1}, k_t) = N[w_t; F(k_t)w_{t-1} + u(k_t), \sigma^2], \quad t = 2, \ldots, T, \tag{2}$$

where $F(1) = 1$, $F(2) = 0.95$, $F(3) = 0.95$, $F(4) = 1$, and $u(1) = 0$, $u(2) = 0.15$, $u(3) = -0.15$, $u(4) = 0$ for each of the four categorical classes of geologic environment, and where $f(w_1|k_1) = N(w_1; 0, \sigma_0^2)$. The expectation terms in the Gaussian Markov model in equation (2) reflect the assumed changes in clay content conditional on the geologic environment. For example, if we are in a fining upwards state, $k_t = 2$, we expect the logit clay content to increase, hence $u(2) = 0.15$, but we include $F(2) = 0.95$ to prevent the logit clay content from exploding.

2.3 DISCRETE MARKOV CHAIN FOR GEOLOGIC ENVIRONMENT

The geologic variable $k_t \in \{1, 2, 3, 4\}$, $t = 1, \ldots, T$ is modeled by Markov transitions:

$$Pr(k_t = j|k_{t-1} = i) = p_{ij}, \qquad t = 2, \ldots, T, \tag{3}$$

where k_1 is fixed, and where $p_{ij} \geq 0$, $\sum_{j=1}^{4} p_{ij} = 1$, for all i. We denote the probability distribution for the geologic sequence conditional on the Markov transition probabilities by $f(k|p)$.

The transition probabilities of the discrete Markov chain are assigned Dirichlet pdfs for each row of the transition matrix p. This pdf is denoted $f(p)$. In this pdf we impose our prior belief about the geologic alternation styles which are; blocky sands followed by a fining upward sequence which ends in blocky shales, thereafter a coarsening upwards sequence which ends in a blocky sands, and so on.

2.4 SAMPLING FROM THE POSTERIOR DISTRIBUTION

The posterior for the hidden variables is defined by

$$f(p, k, w|y) \propto f(y|w)f(w|k)f(k|p)f(p). \tag{4}$$

The hierarchical model assumes conditional independence: For example $f(w|k, p) = f(w|k)$. This simplifies the modeling and simulation, but could of course lead to undesired scenarios in some cases. It is not possible to assess the posterior in equation (4) analytically. We choose to explore the distribution by a Gibbs sampler similar to the one in Carter and Kohn (1996):

1. Intialize (p^1, k^1)
2. Iterate, for $l = 1, \ldots, L$:

   – Draw the rows of the transition matrix from conjugate Dirichlet pdfs. This is a sample from $f(p^{l+1}|k^l)$.

– Draw geologic sequence using the forward and backward loop proposed in Carter and Kohn (1996). This is a sample from $f(k^{l+1}|p^{l+1}, k^l, y)$.

As part of the sampling step for geologic sequence we use an extended Kalman filter to marginalize over values of logit clay content.

3 Examples

We present two examples of the modeling procedures; one using synthetic data generated from the the V-shaped model, and the other based on real well log data.

3.1 SYNTHETIC MODELING

In this synthetic case we generate velocity and density data according to the hierarchy in Figure 1. Figure 2 (left) shows the synthetic log of geologic variables, starting in blocky shales at depth 100, and the corresponding (v_P, ρ) logs. Figure 2 (right) displays a crossplot of the observed noisy data together with the theoretical values for the V-shaped trends. Figure 3 (left) shows the proportions of the four different geologic environments as a function of iteration number in the Gibbs sampler described in Section 2.4. The fluctuations in Figure 3 (left) illustrate the mixing in the Gibbs output and the variability we get in proportions for this synthetic case. In Figure 3 (right) we show generated pseudologs of density and velocity obtained by propagating 50 of the realizations for geologic sequence. These pseudologs (Figure 3, right, dots) are plotted on top of the originally simulated logs from Figure 2 (middle), but the original logs are hardly visible because the pseudologs match the original data quite well. Note the bimodal shape in the pseudolog (Figure 3, right) between 0 and 20 which is caused by a misclassification of sand into fining and coaresening in some of the realizations. This is typical for the algorithm in areas where it is hard to discriminate between the different geological classes.

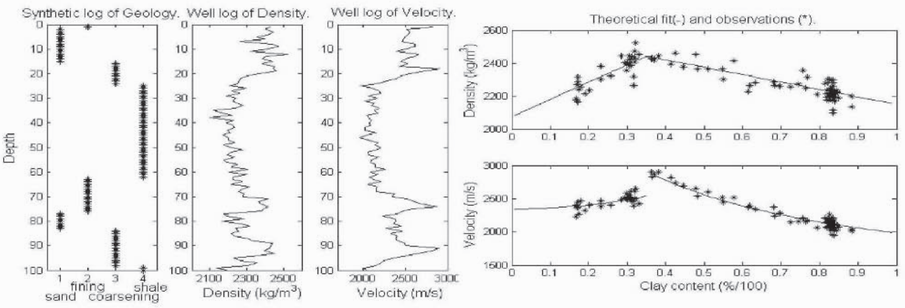


Figure 2. Synthetic logs of geologic sequence with corresponding density and P-wave velocity (left). The observations (right, *) are plotted with the theoretical trend for the V-shaped model (right, solid) as a function of the clay content.

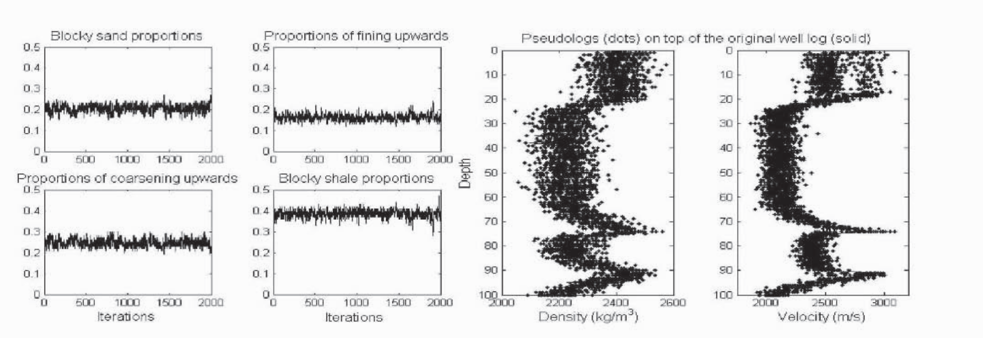


Figure 3. Synthetic case. Left: Proportions of the four geologic classes as a function of iteration number in the Gibbs sampler. Right: Pseudologs of density and velocity (dots). The logs are created by running realizations of geologic sequence through the hierachical model. Original logs of density and velocity (solid).

3.2 WELL LOG DATA FROM A FLUVIAL RESERVOIR

The fluvial well log consists of 600 data points sampled every half foot (15 cm). The observations are plotted versus depth in Figure 4 (left). By using the theoretical properties of quartz, clay and water (Mavko, Mukerji, and Dvorkin, 1998) we assess the parameters required for the V-shaped model in Section 2.1. The results are displayed in Figure 4 (right).

We run the Gibbs sampler described in Section 2.4 for 2000 iterations. Figure 5 (left) shows the gamma ray log as a basis of comparison, Figure 5 (middle) a classified geologic sequence calculated from the Gibbs output. In Figure 5 (left) one part of the gamma ray log is depicted: This zone is what appears to be a coarsening upwards sequence in the gamma ray log near 770 m depth. Our method recognizes the coarsening upwards to some extent since the classification is mostly class 3 near 770 m depth. Figure 5 (right) shows the density and velocity in this part of

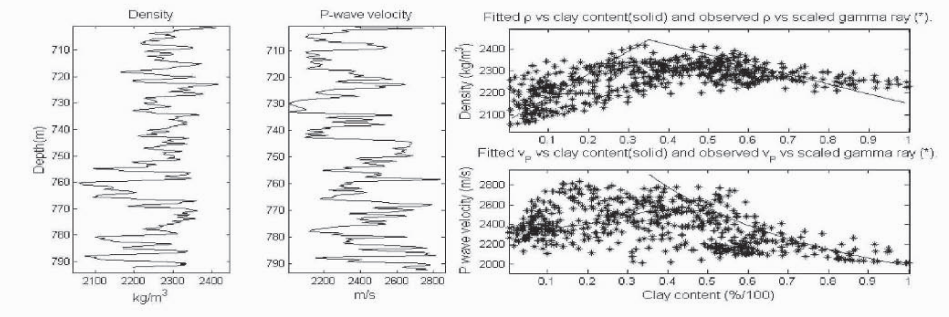


Figure 4. Fluvial case: Density and velocity observations in the well log (left). Fit to the V-shaped model (right), with theoretical fit (solid), well log observations (*).

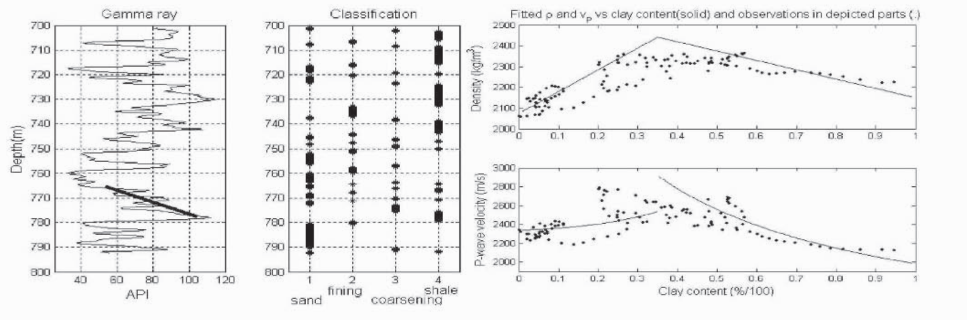


Figure 5. Fluvial case: Gamma ray observations in the well log (left). Classified log of geologic sequence (middle). The depicted area is a coarsening upwards sequence in the gamma ray log which is recognized quite well in the classification. The V-shape for v_P and ρ is quite clear in this depicted zone (right).

the well log, plotted with the theoretically fitted V-shape. The density and velocity observations follow the trend quite nicely in this part of the log. Other parts of the log are harder to classify correctly because the V-shape is missing in the data.

4 Closing Remarks

We propose a hierarchichal model to estimate patterns of shaly-sand lithologic sequences from velocity and density data. The method builds on the sand-shale mix model of Marion et al (1992). When analyzing data from a fluvial well we see that the fit to this model is reasonable in selected sub-sequences of shaly-sands that span from sands to shales through shale-sand mix. The fit is not so good when there are direct sand-shale jumps. Possible extensions include the model by Koltemann and Gorelick (1995), and to use more discrete facies classes to account for jumps. The parameters of the statistical model also have to be more carefully specified.

References

Carter, C.K., and Kohn, R., *Markov chain Monte Carlo in Conditionally Gaussian State Space Models*, Biometrika, v. 83, no. x, 1996, p. 589-601

Eidsvik, J., Mukerji, T., and Switzer, P., *Estimation of geological attributes from a well log: An application of hidden Markov chains*, Mathematical Geology, v. 69, no. 4, 2004, p. 379-397

Fearnhead, P., and Clifford, P., *On-line inference for hidden Markov models via particle filters*, Journal of Royal Statistical Society, Series B, v. 65, no. 4, 2003, p. 887-899

Koltermann, C.E., and Gorelick, S.M., *Fractional packing model for hydraulic conductivity derived from sediment mixtures*, Water Resources Research, v. 31, no. 12, 1995, p. 3283-3297

Marion, D., Nur, A., Yin, H., and Han, D., *Compressional velocity and porosity in sand-clay mixtures*, Geophysics, v. 57, no. 4, 1992, p. 554-563

Mavko, G., Mukerji, T., and Dvorkin, J., *The Rock Physics Handbook*, Cambridge, 1998

Weissman, G.S., and Fogg, G.E., *Multi-scale alluvial fan heterogeneity modeled with transition probability geostatistics in a sequence stratigraphy framework*, Jorunal of Hydrology, v. 226, no. 1-2, 1995, p. 48-65

EVALUATION OF STOCHASTIC EARTH MODEL WORKFLOWS, VERTICAL UP-SCALING AND AREAL UP-SCALING USING DATA FROM THE EUNICE MONUMENT SOUTH UNIT (NEW MEXICO) AND THE LL-652 CENTRAL FAULT BLOCK (VENEZUELA) RESERVOIRS

W. SCOTT MEDDAUGH
ChevronTexaco Energy Technology Company, Bellaire, TX

Abstract. Earth models generated for use in fluid flow simulation incorporate varying amounts of geological detail. The increased costs associated with a detailed earth model are worthwhile if the uncertainty of reservoir performance predictions is significantly reduced. Using data from the Eunice Monument South Unit (EMSU) reservoir and the LL-652 Central Fault Block (CFB) reservoir, workflows that incorporate varying amount of geological information were evaluated on the basis of fluid flow characteristics. Multiple realizations were up-scaled and flow characteristics evaluated using 3D streamlines. There is little difference between the modeling workflows in terms of overall fluid flow characteristics. While the similarity in ultimate recovery was expected the small difference in the oil production rates and water injection rates for the different workflows was not expected. Based on data only from the EMSU models, vertical scale-up has little effect on fluid flow characteristics compared to areal up-scaling. As areal scale-up increases reservoir flow become significantly more "optimistic" as predictions typically overestimate the recovery for a given amount of injection. These results suggest that the more critical "geological" issue in up-scaling may be the areal dimension rather than the vertical dimension.

1 Eunice Monument South Unit (EMSU) Reservoir – New Mexico

The Eunice Monument South Unit Field (EMSU) is located in Lea County, New Mexico. It is largely a stratigraphic play and produces from the Permian (Guadalupian) Grayburg Fm. at an average depth of approximately 3900 feet. The productive interval averages about 250 feet thick and is largely dolomite. The field was discovered in 1929. The OOIP is about 1000 MMSTB of which about 135 MMSTB has been produced. The field currently is under waterflood and produces around 2800 BOPD from over 250 wells. Within the study area the porosity by stratigraphic layer averages 7-9% and permeability averages 0.5-3 md.

The reservoir consists of porous and permeable dolomitized grainstones and mud-poor packstones deposited in high-energy shelf-crest shoals. These shoals were deposited on a carbonate ramp as a series of shallowing-upward fifth order cycles consisting of dominantly carbonate and minor siliciclastic sediments. Updip are non-porous wackestones and mudstones of tidal flat origin that form the lateral updip seal. Downdip

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 743–749.

of the shoal a deeper-water slope setting is inferred. The reservoir is divided into eight layers (only the upper four layers are included in this study) that are separated by thin, very fine-grained sandstone layers. These sandstone layers were initially deposited by aeolian processes during low-stands in sea level and were reworked during flooding at sea level rise. These sandstone layers are well developed over most of the field and are useful in dividing the reservoir into the eight stratigraphic layers (Lindsay, 2000, Meddaugh and Garber, 2000).

2 LL-652 Reservoir (Lagunillas Field) – Venezuela

This reservoir is located in the east flank of the LL-652 anticline (Lagunillas Field) and covers an area of approximately 7,800 acres. In this reservoir oil and gas are produced from some of the most prolific successions in the Maracaibo basin, namely the Eocene C-3-X / C-4-X members of the Misoa Formation. Reservoir production began in 1954. Initial OOIP for the C4.X.01 reservoir is about 640 MMSTB. Cumulative oil production to date is about 84 MMSTB (13.4% OOIP). Current production is under 4000 BOPD. The average porosity ranges from 4% in non-reservoir layers to about 14% in the high quality reservoir layers. The average permeability ranges from less than 0.5 md in non-reservoir layers to 40 md in the high quality reservoir layers. The overall reservoir net to gross ratio is about 0.3 (Moros-Leon and Meddaugh, 2004).

The studied intervals are part of the large scale transgressive-regressive stratigraphic cycle that characterizes the Misoa Formation. This large scale cycle is punctuated by smaller-scale sequences witch represent higher frequency regressive-transgressive cycles within the succession. Within these smaller sequences, successive lowstand (LST), transgressive (TST) and highstand (HST) systems track can be recognized and correlated throughout the entire LL-652 Area. This sequence hierarchy has formed the basis for a sequence stratigraphic subdivision of the C-4-X.01 reservoir. The resulting vertical succession consist of three low-order deltaic units formed by the regular alternation of sandy delta plain sediments with marine mud and sand-prone deposits that prograded northeast as the basin subsided (Moros-Leon and Meddaugh, 2004).

Petrophysical facies characterization indicates that depositional facies exert the primary control over reservoir properties. Reservoir flow units are associated with the amalgamated distributary channels and tidal sand bars that define the LST of each sequence. Permeability baffles and barriers occur within the tidally-influenced TST and HST deposits (Moros-Leon and Meddaugh, 2004).

3 EMSU Workflows

The EMSU reservoir models were generated using the following three workflows with increasing geological constraints:

1. **Simple** – Starting with structural surface for the top of Grayburg Fm and the well picks for the upper four stratigraphic layers provided by the operating company (Opco), five tied surfaces were generated to provided the model framework. Following stratigraphic grid definition (25' areal cell size, 400 layers about 1' thick)

and geometry initialization, the stratigraphic grids were populated with porosity by SGS using the layer appropriate porosity data and semivariograms. Fifteen porosity realizations were generated. Permeability was added to each realization using a transform equation.

2. **Facies** – Using the framework of the simple workflow model the stochastic distribution of the shoal and lagoon facies was done using SIS with a layer-specific facies map as an additional soft constraint. Five facies realizations were generated. Porosity based only on the wells with core data (and hence accurate facies data) was next added using the SGS algorithm constrained by facies region and stratigraphic layer. Permeability was added using the facies-specific transforms.
3. **Lithology-based** – Using the same framework of the simple workflow model the stochastic distribution of the lithology was done using the multi-binary SIS algorithm. Three lithology realizations were generated for each of the five facies realizations generated by the facies-based workflow. Porosity based only on the wells with core data (and hence accurate lithology data) was next added using the SGS algorithm constrained by lithology region and stratigraphic layer. Permeability was added using the lithology-specific transforms.

Additional information on the stochastic modeling workflows for the EMSU models is given by Meddaugh and Garber (2000).

The realizations were scaled-up via a flux-based algorithm to 18 layer models with 50' areal cell size (initial models use a 25' areal cell size) and the flow characteristics evaluated using 3D streamline-based simulator. The stratigraphic framework is maintained during up-scaling. The streamline model used to evaluate the fluid flow characteristics of the three earth model workflows is based on then current field practice (Villegas et al., 1999). The model used five-spot waterflood patterns with production and injection wells spaced approximately 1400' apart. Producers were assigned minimum flowing bottomhole pressures of 500 psia. Injectors were assigned maximum bottomhole injection pressures of 3500 psia. The wells were assigned pressure rather than rate constraints to allow the permeability distributions to drive the results. Note that the choice to up-scale to 18 vertical layers was based on the fact that 18 layers were used for this interval in the full-field model used by Villegas et al. (1999).

4 LL-652 Workflows

Reservoir models were generated using four different workflows with increasing geological constraints:

1. **Simple** – framework from surfaces for the five major stratigraphic picks and maps provided by the Opco. SGS without any region-based constraint was used to distribute effective porosity. Finally, permeability was added via collocated cokriging using effective porosity as soft data. This is the fastest and least expensive method (time and data acquisition/analysis).
2. **RnR (Reservoir/non-Reservoir)** – framework from Simple model. Reservoir and non-reservoir facies distributed using SIS. SGS by facies region was next used to

distribute effective porosity. Last, permeability was added by collocated cokriging with SGS (CCK-SGS) using effective porosity as soft data.

3. **Facies** – framework from Simple model. Seven lithofacies facies distributed using multi-binary SIS (MBSIS). The lithofacies data information was derived from well log signature analysis and three cored wells. Next, SGS by lithofacies region was used to distribute effective porosity. Last, permeability was added by CCK-SGS using effective porosity as soft data.
4. **Complex Facies** – framework from surfaces for 16 "detailed" sequence stratigraphic picks (includes the five main horizons used for the models listed above). Next, seven lithofacies (cross bedded sandstone (SS), tidal delta SS, sheet SS, heterolithic fill, channel shale (Sh), tidal delta Sh, and Transgressive Sh) were distributed using MBSIS by stratigraphic layer. The lithofacies data information was derived from well log signature analysis and three cored wells. Next, SGS by lithofacies region was used to distribute effective porosity. Last, permeability was added by CCK-SGS using effective porosity as soft data.

Additional information on the stochastic modeling workflows for the LL-652 models is given by Meddaugh et al. (2004).

The realizations were scaled-up via a flux-based algorithm to 20 layer models with 50' areal cell size (initial models use a 25' areal cell size) and flow characteristics evaluated using 3D streamline-based simulator. The stratigraphic framework is maintained during up-scaling. The streamline models used a 5-spot waterflood pattern, with production and injection wells spaced approximately 410m apart, respectively. Producers were assigned minimum flowing bottomhole pressures of 3200 psia, and injectors were assigned maximum bottomhole injection pressures of 3700 psia. The wells were assigned pressure rather than rate constraints to allow the permeability distributions to drive the results. Note that the choice to up-scale to 20 vertical layers was based on the fact that 20 layers were used for this interval in the full-field model used by Todd et al. (2002).

5 EMSU and LL-652 Streamline Simulation Results and Conclusions

5.1 FULL FIELD RESULTS

The 3D streamline results (oil and water rate vs. time and HCPVI, oil and water cumulative vs. time and HCPVI, water injection rate vs. time and HCPVI, and cumulative water injection vs. time and HCPVI) obtained from the three EMSU and four LL-652 workflow models show little, if any, difference suggesting that the added geological data provides little, if any, value on a field-wide basis. The table below summarizes the recovery factor (RF) at the end of run (8 years) for the various models. Individual well pairs were not analyzed as that was not the focus of the study. It is likely that significant differences in fluid flow behavior between individual well pairs do exist.

EMSU	Workflow Case	Mean RF	Std Dev RF	Min RF	Max RF	Range RF	Runs
	Base	0.311	0.006	0.303	0.324	0.021	15
	Facies	0.294	0.010	0.274	0.308	0.034	15
	Lithology	0.315	0.008	0.302	0.338	0.036	15
LL-652							
	Base	0.453	0.007	0.442	0.473	0.031	15
	RnR	0.456	0.002	0.452	0.461	0.009	15
	Facies	0.440	0.009	0.418	0.450	0.032	15
	Complex	0.454	0.007	0.436	0.464	0.028	15

Table 1. Summary of 3D streamline recovery factor (RF) results from models derived from the various workflows.

The primary reasons that neither data set yielded models with different flow characteristics among the examined workflows are: (1) the porosity vs. permeability relationship for the various facies and lithologies in both data sets are quite similar (e.g. for EMSU, there was little difference in porosity vs. permeability crossplots of the shoal or lagoon facies or in porosity vs. permeability crossplots of the grainstone or wackestone lithology-types) and (2) the abundance of well control.

Figure 1 summarizes the practical implications of the study results. Note that there is no recommendation made as to value of detailed stratigraphic studies in general as there are numerous opportunities in a field's production history when such studies are critical. However, for the specific cases studied (e.g. abundant well data and little difference in the porosity vs. permeability trends of the various lithofacies) there is little real value added by doing detailed work if one is only interested in a field-level fluid flow understanding.

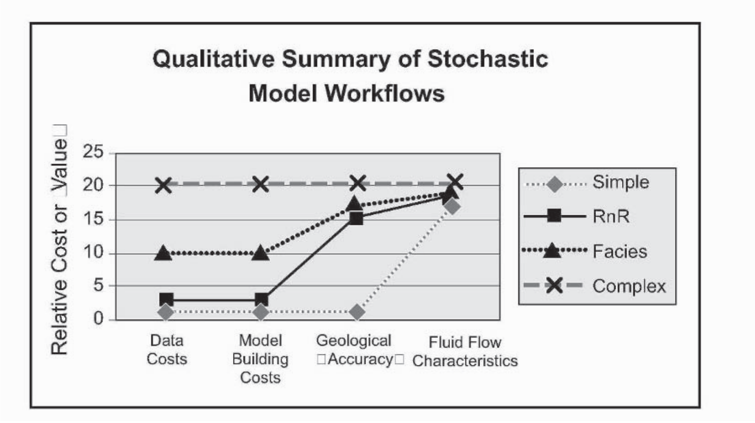


Figure 1. Qualitative comparison of "cost" and "value" of the various workflows examined in this study.

5.2 VERTICAL AND AREAL SCALE-UP RESULTS USING EMSU MODELS

Vertical scale-up has little or no effect on fluid flow characteristics (rates or cumulatives vs. time or HCPVI) over the range studied (400 > 9 layers, scale-up factor 25% > 0.5%). The RF for the various levels of vertical scale-up are summarized in Table 2. These results suggest that the critical "geological" issue in scale-up is not the vertical dimension although clearly there must be a point at which vertical scale-up does create a "tank-like" reservoir with average rock properties.

Upscale Factor	Number of Layers	Mean RF	Std Dev RF	Min RF	Max RF	Range RF	Runs
25%	400	0.320	na	na	na	na	1
10%	167	0.321	na	na	na	na	1
5%	83	0.320	0.006	0.311	0.332	0.020	15
1%	18	0.311	0.006	0.303	0.324	0.021	15
0.10%	9	0.328	0.006	0.320	0.339	0.018	15

Table 2. Recovery Factors (RF) for Varying Levels of Vertical Up-scaling.

Areal scale-up has significant effects on fluid flow characteristics. As real scale-up increases reservoir flow characteristics become more optimistic (less water injected, lower rate for same oil recovery). This may have a significant impact on the number of cells needed between wells in finite difference fluid flow simulation. This study suggests that 3-5 cells between well is not sufficient. More study is needed to fully assess areal scale-up issues. The RF for the various levels of areal scale-up are summarized in Table 3. Note that RF decreases significantly as model cell size after up-scaling is increased (model becomes more "tank-like").

Areal Cell Size (feet)	Simple Mean RF	Std Dev RF	Min RF	Max RF	Range RF	Runs
25.00	0.339	0.005	0.331	0.348	0.017	15
50.00	0.311	0.006	0.303	0.324	0.021	15
100.00	0.307	0.005	0.299	0.318	0.018	15
150.00	0.302	0.006	0.291	0.314	0.023	15
Areal Cell Size (feet)	**Lithology Mean RF**	**Std Dev RF**	**Min RF**	**Max RF**	**Range RF**	**Runs**
25.00	0.349	0.006	0.339	0.361	0.023	15
50.00	0.315	0.008	0.302	0.338	0.036	15
100.00	0.306	0.006	0.296	0.320	0.024	15
150.00	0.296	0.007	0.285	0.312	0.027	15

Table 3. Recovery Factors (RF) for Varying Levels of Areal Up-scaling.

This page intentionally blank

Conclusions

This study does not argue "against" doing detailed geological and stratigraphic analyses. This study does, however, argue the case that it is possible to put too much "geology" in earth models destined for fluid simulation. Unfortunately, this study does not provide quantitative guidelines for what constitutes "too much" geology except that if the porosity vs. permeability relationship is essentially identical within the geological, stratigraphic, facies, or lithology "containers" that could be used as model constraints there is little to be added by using such containers as additional model constraints. This study also suggests that additional work is needed to define "optimum" areal cell sizes when fine-scaled geological models are up-scaled for fluid flow simulation.

References

Lindsay, R.F. et al, 2000. *Role of Sequence Stratigraphy in Reservoir Characterization: An Example from the Grayburg Formation, Permian Basin in Permian Basin Exploration and Production Strategies*, West Texas Geological Symposium, November 5-6, 1992, p. 19-26.

Meddaugh, W.S. and Moros, S.J., 2004. *Facies, Sequence Stratigraphy and Flow Units in the Eocene C-4-X.01 Reservoir, LL-652 Area (Lagunillas Field), Western Venezuela*, AAPG (Dallas, April).

Meddaugh, W.S., Moros, S.J., and Olivier, W., 2004. *Stochastic Reservoir Modeling and Workflow Evaluation, C3-C4 Interval, Central Fault Block (CFB), LL-652 Reservoir, Venezuela*, AAPG (Dallas, April).

Meddaugh, W.S. and Garber, R.A., 2000. *Reservoir Characterization and Geostatistical Modeling of the Eunice Monument South Unit Field*, New Mexico, AAPG Hedberg Conference.

Todd, W. W. et al., 2002. LL652 Field Central Fault Block Full Field Simulation Study, EPTC, Houston (internal company report).

Villegas, M. E., et al., 1998. *Eunice Monument South Unit Full Field Reservoir Simulation Study* – 1998, CPTC, Houston, #TM9900125 (internal company report).

APPLICATION OF DESIGN OF EXPERIMENTS TO EXPEDITE PROBABILISTIC ASSESSMENT OF RESERVOIR HYDROCARBON VOLUMES (OOIP)

W. SCOTT MEDDAUGH, STEWART D. GRIEST, STEPHEN J. GROSS
ChevronTexaco Energy Technology Company, Bellaire, TX

Abstract. Design of Experiment (DoE) methodology was used to minimize the number of stochastic earth models that were needed to appropriately evaluate original oil in place (OOIP) uncertainty for a Jurassic-age, Middle East carbonate reservoir. The DoE methodology enables the maximum amount of information to be obtained from the minimum number of experiments (model OOIP) in which multiple parameters (structural uncertainty, facies distribution uncertainty, oil-water contact uncertainty, net-to-gross uncertainty, etc.) contribute. The DoE methodology also allows for rapid determination of the magnitude of model parameters to overall OOIP uncertainty. Thus, attention can properly be focused on the few key model parameters that most affect OOIP uncertainty, perhaps to the point of obtaining additional data if cost-justified.

The DoE-based workflow used was as follows: (1) use Plackett-Burman design (one of several DoE methodologies tested) to determine which combinations of model parameters should be evaluated; (2) collect the experimental results (OOIP); (3) analyze the results statistically to determine significant contributors to OOIP uncertainty; (4) use the experimental results to obtain a response "surface" (equation) that describes the relationship between OOIP and the significant contributors to OOIP uncertainty; (5) use the response surface along with appropriate statistical distributions for the significant contributors to OOIP uncertainty in a Monte Carlo-process to obtain P_{10}, P_{50}, and P_{90} OOIP values. Drained volume uncertainty was also evaluated using the above workflow so that stochastic reservoir models with P_{10}, P_{50}, and P_{90} drained volumes could be generated using appropriate combinations of geologically reasonable parameters for further sensitivity and optimization studies as well as input to probabilistic economic evaluation.

1 Introduction

Design of experiments (DoE), also often referred to as experimental design (ED), is seeing increasing use within the oil and gas industry within both the reservoir geology and reservoir engineering communities (Friedmann et al, 2003; White and Royer, 2003; Peng and Gupta, 2003; Sanhi, 2003, White et al., 2001; Peng and Gupta, 2004). Within the reservoir engineering community, DoE techniques are now routinely used in reservoir fluid flow simulation studies to reduce the number of simulation sensitivity or

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 751–756.

optimization runs. Within the reservoir geology community, DoE techniques are being used, though not yet routinely, to assess reservoir uncertainty – both volumetric (OOIP) and connectivity (drained volume) uncertainty.

The focus of this short communication is to examine the use of DoE to assess reservoir uncertainty – both volumetric and connectivity – using data from a Jurassic-age, Middle East carbonate reservoir. A DoE-based methodology was employed for this study to allow efficient and quantitative examination of both OOIP and drained volume uncertainty so that P_{10}, P_{50}, and P_{90} reservoir models could be developed for use in additional reservoir sensitivity studies, field development optimization studies, and economic evaluation.

The methodology of this project was based on the Plackett-Burman (PB) experimental design, which although generally regarded as a screening design is the most efficient two level design (Plackett and Burman, 1946) and ideally suited to the short time frame for the evaluation of the Jurassic-age, Middle East carbonate reservoir. Once the uncertainty factors and levels were determined, the DoE-based workflow followed for the Jurassic-age, Middle East carbonate reservoir study was as follows: (1) use the Plackett-Burman design to determine which combinations of model parameters should be evaluated; (2) collect the experimental results (OOIP); (3) analyze the results statistically to determine significant contributors to OOIP uncertainty; (4) use the experimental results to obtain a response "surface" (equation) that describes the relationship between OOIP and the significant contributors to OOIP uncertainty; (5) use the response surface along with appropriate statistical distributions for the significant contributors to OOIP uncertainty in a Monte Carlo-process to obtain P_{10}, P_{50}, and P_{90} OOIP values. Drained volume uncertainty was also evaluated using the above workflow so that stochastic reservoir models with P_{10}, P_{50}, and P_{90} drained volumes could be generated using appropriate combinations of geologically reasonable parameters for further sensitivity and optimization studies as well as input to probabilistic economic evaluation.

2 Middle East Example

The Jurassic-age, Middle East carbonate reservoir that is the primary focus of this note is located largely within the Partitioned Neutral Zone (PNZ) between Saudi Arabia and Kuwait. The reservoir was discovered in 1998 and currently produces 25-33º API oil with 5% water cut from five wells. The reservoir depth is about 9700'. The reservoir, a relatively simple four-way closed anticline oriented NW-SE, produces largely from limestone within the Marrat interval. The porosity in productive zones averages about 12%. Permeability is low (average for interval is on the order of 1-5 md) but extremely variable with measured core plug values up to 400 md. Well test derived permeabilities range between 10 and 80 md. Although a detailed sequence stratigraphy for the interval has not been finalized, there is good correlation between all five wells. The available data suggests that porosity increases significantly updip and is almost certainly related to an up-dip facies change as is shown in table below (wells 4, 5, and 8 are structurally low compared to wells 6 and 7).

Porosity (%)	**All Wells**	**W-4**	**W-5**	**W-8**	**W-6**	**W-7**
Zone A	6.6	5.3	5.0	5.2	8.8	8.6
Zone C	7.6	5.9	8.6	7.2	9.9	8.1
Zone E	9.8	8.1	8.3	7.4	11.6	12.8

The average well log water saturation (Sw) for Zone A is 45%, for Zone C it is 51%, and for Zone E it is 31%. OOIP and drained volume calculations detailed later in this note make use of a normalized J-function derived Sw rather than the well log calculated Sw. The J-function derived Sw values are in very good agreement with the well log Sw values in the productive portions of the overall reservoir interval. Thus, OOIP as calculated from well log Sw values and OOIP calculated from J-function derived Sw values are also in very good agreement. Porosity (and Sw) histogram uncertainty was estimated based on well to well variability.

The original oil water contact (OOWC) is unknown. For the upper reservoir zones, the oil water contact is defined by a lowest known oil (LKO) at -9329' and a lowest closed contour (LCC) at -9909'. The lower zone may have a separate OOWC with a LKO at -9831' and a LCC at -9909'.

Production data and well tests strongly suggest that some intervals, notably in the upper portion of the reservoir, may be fractured although image log data show relatively few fractures. 3D seismic interpretation shows some faulting likely within the interval, although the resolution of the seismic volume does not permit easy identification and/or mapping of the faults. One well (W-5) has significantly lower gravity oil that may reflect some reservoir compartmentalization.

The table below summarizes the factors were considered to impact both OOIP and drained volume uncertainty. The high and low values for each factor were chosen to represent likely P_1 and P_{99} scenarios.

Uncertainty Factor	**Low Drained Volume (low OOIP) Case**	**Mid Drained Volume (mid OOIP) Case**	**High Drained Volume (high OOIP) Case**
Structure	Current structure map minus uncertainty map A (tied to all wells)	Per current structure map (tied to all wells)	Current structure map + uncertainty map B (tied to all wells)
Facies	No facies	One facies with moderate porosity improvement defined by wells W-6 and W-7.	Two facies with significant (well W-7) and moderate porosity improvement (well W-6).
Porosity Histogram	Well data with -2 porosity unit (p.u.) shift	Given by well data	Well data with +2 p.u. shift
Sw Histogram	J-function with +10 saturation unit (s.u.). shift	Given by J-function	J-function with -10 s.u. shift
OOWC	-9329' for A, C zones; -9831' for E zone	-9500' for A, C zones; -9850' for E zone	-9909'

Uncertainty Factor (continued)	**Low Drained Volume (low OOIP) Case**	**Mid Drained Volume (mid OOIP) Case**	**High Drained Volume (high OOIP) Case**
Porosity vs. Sw Correlation	Not used as Sw derived via J-function	Not used as Sw derived via J-function	Not used as Sw derived via J-function
Porosity Semivariogram R1	500 m	1500 m	6000 m
Permeability Multiplier	None. Permeability distribution given by core data and porosity to permeability transform(s)	Guided by global multiplier needed for prior fluid flow simulation study (e.g. 6.5x)	Guided by well test derived permeability (e.g. x 10)
Faults	Two per current structure map	Four faults parallel to structure with three "perpendicular" faults defining a total of 6 "compartments"	Four faults parallel to structure with eight "perpendicular" faults defining a total of 12 "compartments"
Fault Transmissibility	Sealing (transmissibility = 0.005)	Moderate (transmissibility = 0.05)	Essentially open (transmissibility = 0.5)

The structural uncertainty incorporated in the analysis essentially has little uncertainty near the five wells (+20', -10') and increases to a maximum value (+250', -125') 2 km from the wells. Porosity and Sw uncertainty ranges set based on the range of individual well average values (table given previously). The porosity semivariogram range uncertainty is based on analog carbonate reservoirs. Likewise, the facies distribution uncertainty is based on analog carbonate reservoirs.

The thirteen reservoir model scenarios given by the Plackett-Burman design (PB) design table were generated using the following workflow:

1. Build structural framework and stratigraphic model grids for minimum, mid, and maximum uncertainty cases. All stratigraphic model grid top and bottom surfaces are tied to the appropriate well picks
2. Distribute porosity using sequential Gaussian simulation by stratigraphic layer using layer appropriate histograms and semivariogram ranges per the experimental design table. As appropriate, modify the porosity distribution (histogram) per the experimental design table.
3. Distribute minimum case permeability using the porosity to permeability transforms given previously. As appropriate, modify the permeability distribution (histogram) per the experimental design table.
4. Distribute Sw using J-function. As appropriate, modify the Sw distribution (histogram) per the experimental design table.

The calculated OOIP for each model was then calculated and the results analyzed statistically to determine which factors significantly affect OOIP uncertainty. The analysis showed that OOWC and the porosity histogram were clearly the most significant uncertainty sources (Figure 1). The Sw histogram and structural uncertainty were next most important. Other factors were not important based on a 95% confidence limit. Clearly, if our OOIP assessment is to be improved, a better understanding of the

OOWC and porosity histogram is critical. Such information can have a value assigned and be used as part of the decision process for the next appraisal well.

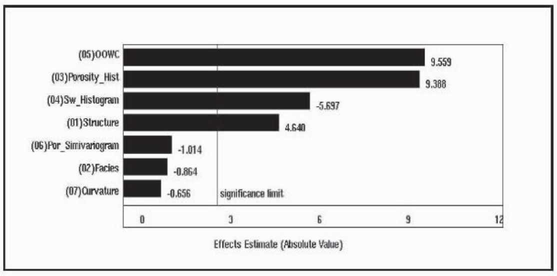


Figure 1. *Pareto chart showing relative contribution of each uncertainty source relative to OOIP uncertainty. Significance limit shown corresponds to 95% confidence limit.*

As noted above, another significant source of uncertainty in reservoir management is connectivity. Connectivity uncertainty was evaluated for the Jurassic-age, Middle East carbonate reservoir by evaluating the drained volume by finite difference simulation on an up-scaled static model from each of the 13 scenarios given in PB design table. Finite difference simulation was selected because model run times were very short. For larger models streamline-based simulation may be more appropriate. The following producing rules were used to evaluate drained volume:

- Well spacing = 160 acres
- Start date = 1 January 2006
- End date = 1 January 2036
- Maximum liquid rate = 3000 BOPD/well
- Minimum bottomhole pressure = 750 psia
- Economic limit oil rate = 100 BOPD
- Economic limit water cut = 80%

Statistical analysis of the drained volume results showed that the only significant sources of uncertainty at a 95% significance level were the porosity histogram and the permeability multiplier (Figure 2). All other factors including structural, OOWC, and faulting uncertainty were not significant. These results, together with the derived response surface equation were used to build the P_{10}, P_{50}, P_{90} drained volume earth models for subsequent sensitivity and optimization studies that will yield P_{10}, P_{50}, and P_{90} flow streams for use in probabilistic economic analysis.

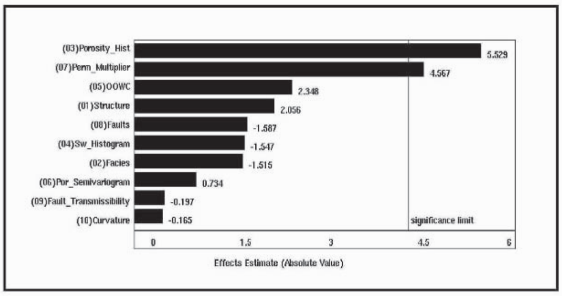


Figure 2. *Pareto chart showing relative contribution of each uncertainty source relative to drained volume uncertainty. Significance limit shown corresponds to 95% confidence limit.*

It should be noted that the emergence of the porosity histogram and the permeability multiplier (which are somewhat linked) as the only significant sources of uncertainty relative to drained volume was unexpected. This "surprise" reinforced the need to use of a DoE-based workflow to assess uncertainty early in reservoir studies. DoE-based workflows, which are scenario-based rather than realization-based, are very efficient and therefore can be used to reduce project cycle time. Such has also been the case for the Jurassic-age, Middle East carbonate project. The scenario-based workflow may require less than 20-25% of the time needed for a "traditional" workflow to provide the same input to an economic analysis. Assessment of uncertainty due to "dynamic" model input parameters (e.g. aquifer support, Kv/Kh, PI multiplier, etc) will be assessed in a second level DoE design that will be completed prior to development optimization and economic modelling.

3 Summary

This study shows the value added of using a DoE-based scenario workflow to assess OOIP and drained volume uncertainty. The value added is largely due to the relatively short cycle time of a DoE-based scenario workflow compared to the cycle time of a realization-based workflow. Additional value of a DoE-based scenario workflow is that even a "cursory" assessment of uncertainty sources early in a project's lifecycle may significantly impact which uncertainty elements are targeted for more extensive study and which uncertainty sources may be "neglected" which oftentimes reduces project cycle time. Uncertainty assessment coupled with a value of information assessment may be sufficient to justify acquisition of additional data.

Acknowledgements

The authors thank the Saudi Arabia Ministry of Petroleum, Saudi Arabian Texaco, and ChevronTexaco for permission to publish this paper.

References

Friedmann, F., Chawathe, A., and LaRue, D.K., 2003. *Assessing Uncertainty in Channelized Reservoirs Using Experimental Designs, SPE Reservoir Evaluation and Engineering*, August 2003, p. 264-274.

Mason, R.L., Gunst, R.F., and Hess, J.L., 2003. *Experimental Design and Analysis of Experiments with Application to Engineering and Science* (2nd Edition), Wiley-Interscience, 728 p.

Peng, C.W. and Gupta, R., 2003. *Experimental Design in Deterministic Modelling: Assessing Significant Uncertainties, SPE 80537* (Indonesia, September, 2003).

Peng, C.W. and Gupta, R., 2004. *Experimental Design and Analysis Methods in Multiple Deterministic Modelling for Quantifying Hydrocarbon In-Place Probability Distribution Curve, SPE 87002* (Kuala Lumpur, March, 2004).

Plackett, R.L., and Burman, J.P., 1946. *The Design of Optimum Mutifactorial Experiments, Biometrika, V. 33. p. 305-325*.

Sanhi, A., 2003. *Case Studies of Uncertainty Analysis in the Seismic to Reservoir Simulation Workflow, SPE 84188* (Denver, October, 2003)

White, C.D. and Royer, S.A., 2003. Experimental *Design as a Framework for Reservoir Studies, SPE 79676* (Houston, February, 2003).

White, C.D., Willis, B.J., Narayanan, K., and Dutton, S.P., 2001. *Identifying and Estimating Significant Geologic Parameters with Experimental Design, SPE Journal*, September 2001, p. 311-324.

STOCHASTIC RESERVOIR MODEL FOR THE FIRST EOCENE RESERVOIR, WAFRA FIELD, PARTITIONED NEUTRAL ZONE (PNZ)

W. SCOTT MEDDAUGH[1], DENNIS DULL[1], STEWART GRIEST[1], PAUL MONTGOMERY[2], and GERRY McNABOE[3]
[1]*ChevronTexaco Energy Technology Company, Bellaire, TX*
[2]*ChevronTexaco Overseas Production Company, Perth, Australia*
[3]*ChevronTexaco North America Upstream Production Co., Bakersfield, CA*

Abstract. The Wafra field is located in the Partitioned Neutral Zone (PNZ) between Saudi Arabia and Kuwait. The field produces from five intervals of which the Tertiary First Eocene dolostone reservoir is the youngest. The reservoir consists of extensively dolomitized peloidal packstones and grainstones that were deposited on a very gently dipping, restricted ramp environment with interbedded evaporites. Discovered in 1954, the First Eocene reservoir has produced more than 280 MMbbls of 17-19° API, high sulfur oil. The stochastic reservoir model utilizes a new sequence stratigraphic framework and is based on data from over 285 wells. The geostatistical model covers a 17.3 x 21.5 km area (372 km^2). Porosity was distributed using sequential Gaussian simulation (SGS) constrained by stratigraphic layer. Porosity semivariogram range parameters average 1500 m (compared to an average well spacing of about 500 m) and show a moderate N120E trend. Permeability was distributed using a cloud transform algorithm that was constrained by core data for specific stratigraphic layers. Water saturation (Sw) was distributed by collocated cokriging with SGS using Sw well log

1. Location and Geological Setting

The First Eocene reservoir is located in Wafra field (Figure 1). The First Eocene is the shallowest of the reservoirs at Wafra field with an average depth of 1000 feet. The First Eocene is about 750 feet thick with an average porosity of 35 percent and an average permeability of 250 md. Oil was first discovered in the First Eocene in 1954. Full scale development and production did not commence until March 1956.

The First Eocene production occurs withindolomitizedpeloidal grainstones and packstones. These rocks were deposited

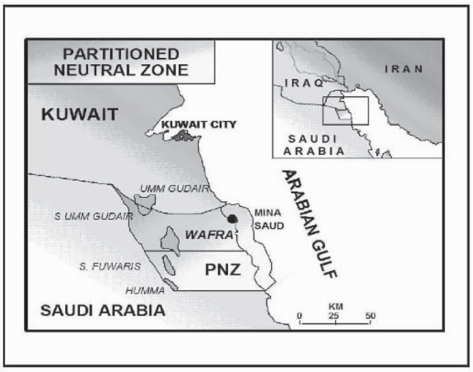


***Figure 1**. Map showing the location of the Wafra field in the Partitioned Neutral Zone (PNZ) between Saudi Arabia and Kuwait.*

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 757–762.

on a gently dipping, shallow restricted ramp under arid to semi-arid conditions. This interpretation is based on the presence of abundant interbedded evaporites and a paucity of normal marine fauna. The abundance of evaporites, primarily in the form of gypsum either present as isolated nodules, coalesced nodules, or bedded, suggests that environmental restriction was sufficient for the development of hyper-saline lagoons or salinas and sabkhas. The Paleocene through the Eocene at Wafra reflects an overall upward shallowing event that culminates in the deposition of the 1st Anhydrite (Rus Formation). The base of the First Eocene reservoir is a succession of anhydrite beds that is locally known as the Second Anhydrite.

2. Sequence Stratigraphy

A new sequence stratigraphic-based correlation scheme was developed for the First Eocene reservoir as part of this study. The sequence stratigraphic framework of the First Eocene was based on five recent cores described with sufficient detail to construct the depositional framework and provide the basis for correlation. The First Eocene framework is based on standard definitions of cycle, cycle set, and high frequency sequence (HFS). A cycle is defined as the smallest set of genetically related lithofacies representing a single base level rise and fall. This is comparable to the parasequence of Van Wagoner et al (1988, 1990). The cycle set is a bundle of cycles that show a consistent trend of aggradation, or progradation (Kerans and Tinker, 1997). The cycle tops for the First Eocene were identified using the following criteria: (1) hardgrounds or erosional surfaces; (2) tidal flat mudstones occasionally with algal laminations; (3) inner shelf peloidal or skeletal dolowackestones (4) bedded or nodular gypsum associated with tidal flat and sabkha deposition; and, (5) dolograinstones and grain-dominated dolopackstones. The latter are mostly confined to the lower portion of the First Eocene in the transgressive systems tract.

Hardgrounds are defined as cemented carbonate rocks that can be encrusted, discolored, bored, rooted, and solution-ridden and are commonly interpreted as representing a gap in sedimentation or an unconformity. In the case of the First Eocene reservoir, intraclastic rudstones occasionally overlay the hardgrounds and sub-aerial exposure surfaces. In some cases brecciation is observed beneath the hardground, indicating intermittent sub-aerial exposure and incipient soil formation in the original limestone.

Figure 2 shows a core description from a First Eocene section from one well in the Wafra field and illustrates the position of features used to identify the cycle caps. The cycles in the First Eocene are primarily defined by drops in base-level that resulted in the development of low-energy depositional conditions. These conditions were conducive to the development of hardgrounds, mud-dominated lithofacies, and evaporite deposition. The exception to this is in the lower portion of the First Eocene where shallowing upward cycles are capped with grain dominated lithofacies reflective of mid-ramp deposition within the transgressive systems tract. The cycle set is generally used for correlation as it represents a thicker stratigraphic interval. The cycle set can be more easily correlated since it will be less impacted by the local topography variations. The grouping of cycles into cycle sets that could be easily correlated was an iterative process using all cored wells and cycle stacking patterns.

The high frequency sequence (HFS) as defined by Kerans and Tinker (1997), is comparable to the composite sequence proposed by Mitchum and Van Wagoner (1991), and is defined as being bounded by a base-level-fall to base-level-rise turnaround. The HFS is composed of genetically related cycles and cycle sets. The bounding surfaces of the HFS are identified by: (1) sub-aerial unconformity; (2) turnaround from progradation to transgression; (3) lithofacies tract offset across a chronostratigraphic boundary (e.g. abrupt change from the deposition of high energy subtidal grainstones to tidal flat facies indicating a significant lowering of base-level); and (4) systematic changes in thickness and/or lithofacies proportion of cycles or cycle sets (e.g. a consistent pattern of upward thinning of cycles or cycle sets that can be correlated as sequence boundaries, reflecting a base-level fall). Based on the available core descriptions at the time of this study the First Eocene has been subdivided into ten interpreted HFS units from the core associated that are bounded by hardgrounds, subaerial exposure surfaces, or lithofacies tract offset. Additional subdivisions below the EOC900 surface are speculative and based on well log data only. The HFS bounding surfaces of the First Eocene are chronostratigraphic.

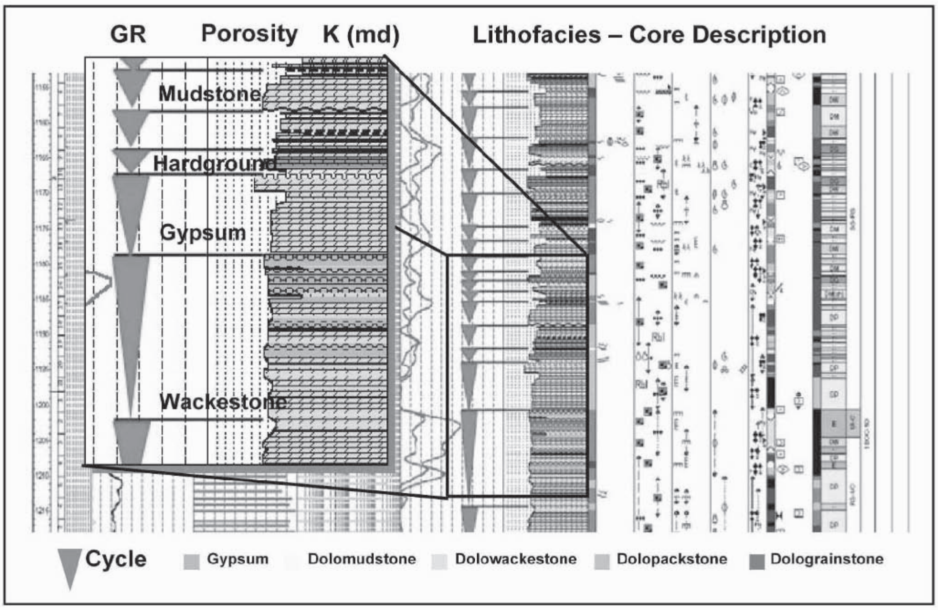


Figure 2. *Detail from core description for a First Eocene well showing the criteria used to distinguish the stratigraphic cycles.*

Correlation of HFS bounding surfaces within the First Eocene section is based on recognition of these surfaces within core as well as by distinctive gamma ray signatures. The shallowing upward cycles are capped by mud-dominated rocks, hardgrounds and exposure surfaces, which are correlative to a higher gamma ray response on the well logs. The gypsum and gypsum-rich tidal flat caps do not show the higher well log gamma ray values, but are most commonly capped with mud-dominated lithofacies. This mud-dominated lithofacies typically has a higher gamma ray signature and helps define the cycle tops on the well logs. Many of the cycle tops and associated well log gamma signature can be correlated across the entire Wafra field. The ability to correlate the fine-scale well log gamma ray pattern and correlative cycles seems to indicate that the First Eocene was deposited as a tectonically stable and mainly aggradational portion of the shelf where subsidence was keeping pace with carbonate deposition. The core description and well log correlation are keys to visualizing the degree of compartmentalization of the First Eocene. The hardgrounds and mud-dominated lithofacies are correlative stratigraphically downdip to more grain-dominated lithofacies. The cycle tops are capped by dolomudstones and hardgrounds which may have low permeability and may be local barriers to flow. The hardgrounds are not regionally correlative, but typically are associated over several kilometers with low porosity, mud-dominated lithofacies.

3. Model Input Data

The average porosity and Sw variation by stratigraphic unit is given in Table 1. Note that below EOC800 there is a general decline in average porosity and a significant increase in average Sw. Note also the local porosity minimum for the EOC500 and EOC600 interval. These interval, which have average porosities near 0.30 have very different average Sw. The average Sw for EOC500 is about 0.70, considerably higher than intervals above or below. The EOC500 interval may act as a field-wide baffle to flow. Core data is available from a total of 10 wells, five of which have high quality core descriptions. Only post-1985 core plug analyses were used. Analyses done prior to 1985 used a high temperature extraction technique during which a substantial portion of the gypsum may be dewatered.

Semivariogram models used for porosity for each stratigraphic layer are summarized in the table below. The models are all exponential form with nugget = 0 and sill = 1.

	Well		**Semivariogram**			
Stratigraphic Interval	Average Porosity	Average Sw	XY Range 1 (meters)	XY Range (meters)	Azimuth	Z Range (meters)
EOC000	0.337	0.715	1600	760	102	1.80
EOC100	0.362	0.541	2040	1260	118	2.04
EOC200	0.359	0.534	1280	700	133	2.02
EOC300	0.352	0.611	1200	650	120	1.89
EOC400	0.357	0.571	1280	1060	113	1.63
EOC500	0.309	0.715	1890	880	131	2.16
EOC600	0.302	0.607	2150	1300	147	2.55
EOC700	0.353	0.621	1280	870	118	1.67
EOC800	0.358	0.732	1240	690	121	2.32
EOC900	0.329	0.822	1520	800	130	3.17
EOC1200	0.309	0.894	1540	1010	115	3.21
EOC1400	0.274	0.965	1180	760	127	3.10
EOC1800	0.261	0.979	1170	690	107	3.00

Table 1. Summary of well log input data and semivariogram models by stratigraphic layer.

4. Stochastic Model

As the First Eocene reservoir lacks an easily defined oil-water contact, the boundaries of the model are arbitrary. The model boundaries were set to include all First Eocene producing wells (past and present) as well as the wells within the southeast "extension" of the field (Figure 3). The model areal grid is 182 x 227 cells (94.8315 m to allow easy calculation of 80 acre pattern results) in areal dimension and includes 535 vertical layers. The vertical layering is nominally one foot between the EOC000 and EOC900 markers and 2-4 feet below EOC900.

The final workflow used to generate the First Eocene earth model was as follows:

1. Build stratigraphic grid framework using relevant top and bottom surfaces for each stratigraphic interval
2. Distribute porosity by sequential Gaussian simulation (SGS) by stratigraphic layer using the layer appropriate semivariogram model and the layer appropriate porosity data. Efforts to incorporate a lithofacies or rock-type constraint were unsuccessful as neither could be reliably predicted using the available well log data.
3. Distribute permeability by cloud transform (minimum of 10 points per bin; maximum of 30 bins) using layer appropriate core porosity-permeability calibration crossplot and prior porosity distribution. Alternative methods including layer specific transforms were evaluated but found to offer no significant advantage over the cloud transform approach. Minor modifications to an almost trivial number of cells were made to reduce the number of cells with anomalously high permeability at low porosity or anomalously low permeability at high porosity.
4. Distribute Sw to model using colocated cokriging with SGS using layer appropriate Sw data from all wells, semivariogram, and correlation coefficient with porosity. The prior porosity distribution was used as the secondary data. The effect of using only data from older, pre-1995 wells (mostly pre-1985) was evaluated and found to be essentially identical to the Sw distribution obtained using all wells.

5. Model Use and Conclusions

The stochastic model generated for the First Eocene reservoir has been used to generate maps to facilitate reservoir management decisions and to calculate probabilistic OOIP. For the Main Area as shown in Figure 3, the P_{50} OOIP of 11.9 x 10^{10} bbls. For the area within well control as shown in Figure 3 the P_{50} OOIP is 24.3 x 10^{10} bbls. The model is also the basis of a full field fluid flow simulation model that is currently being used to evaluate a variety of reservoir management options including in-fill drilling, waterflood EOR, and steamflood EOR. Current analysis shows that steamflood EOR may have the most attractive economics and a small 5-spot (central injector) scale test is planned for late 2005 or early 2006.

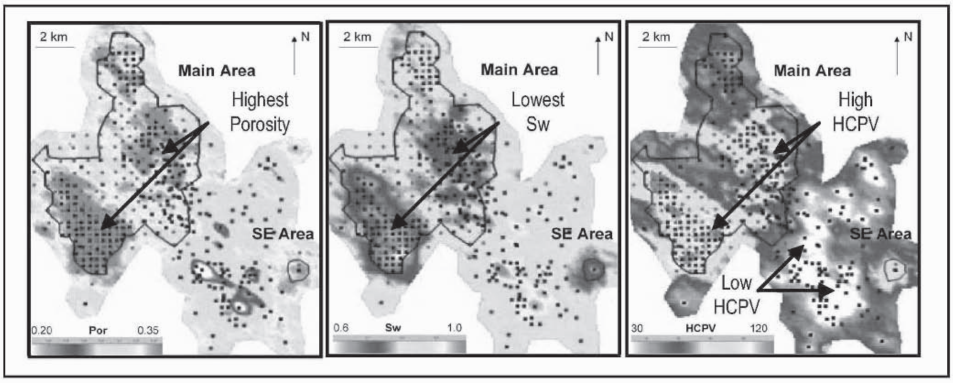


Figure 3. *Average porosity, Sw, and hydrocarbon pore volume (HCPV) maps for the First Eocene reservoir. Polygon shows the well control area as referred* **to in**

Acknowledgments

The authors thank the Saudi Arabia Ministry of Petroleum, Saudi Arabian Texaco, and ChevronTexaco for permission to publish this paper.

References

Kerans, Charles, and W. Scott Tinker, 1997, *Sequence Stratigraphy and Characterization of Carbonate Reservoirs*, SEPM Short Course Notes No. 40.

Mitchum, R. M., and J. C. Van Wagoner, 1991, *High-frequency sequences and their stacking pattern: sequence stratigraphic evidence of high-frequency eustatic cycles, in K. T. Biddle and W. Schlager, eds., The Record of Sea-Level Fluctuations: Sedimentary Geology*, v. 70, p. 131-160.

Van Wagoner, J. C., H. W. Posamentier, R. M. Mitchum, P. R Vail,. Sarg, J. F. Loutit, and J. Hardenbol, 1988, an overview of the fundamentals of sequence stratigraphy and key definitions, in C.K. Wilgus, B. J. Hastings, H. Posamentier, J. C. Van Wagoner, C. A. Ross, and C. G. St. C. Kendall, eds., *Sea-Level Change: An Integrated Approach*, SEPM Special Publication no. 42, p. 39-46.

Van Wagoner, J. C., R. M. Mitchum, K. M. Campion, and V. D. Rahmanian, 1990, *Siliclastic sequence stratigraphy in well logs, cores, and outcrops: concepts for high-resolution correlation of time and facies: AAPG, Methods in Exploration Series*, no. 7, p. 55.

MULTIPLE-POINT STATISTICS TO GENERATE PORE SPACE IMAGES

HIROSHI OKABE[1, 2] and MARTIN J. BLUNT[1]
[1] *Department of Earth Science & Engineering, Imperial College London, SW7 2AZ, UK*
[2] *Japan Oil, Gas and Metals National Corporation, 1-2-2 Hamada, Mihama-ku, Chiba-shi, Chiba, 261-0025, Japan*

Abstract. Multiple-point statistics (MPS) on a two-dimensional (2D) thin section image are used to generate a three-dimensional (3D) pore space image with an assumption of isotropy for orthogonal planes. The method gives images that preserve typical patterns of the void space seen in thin section. Using only single and two-point statistics in the reconstruction often underestimates the void connectivity, especially for low porosity materials; however, multiple-point statistics method significantly improves the void connectivity. The method is tested on sandstone and carbonate samples. Permeability is predicted directly on the 3D images using the lattice Boltzmann method (LBM). The numerically estimated results are in good agreement with experimentally measured permeability. Furthermore, the method provides an important input for the creation of geologically realistic networks for pore-scale modeling to predict multiphase flow properties.

1 Introduction

The reconstruction of 3D porous media is of great interest in a wide variety of fields, including earth science and engineering, biology, and medicine. Several methods have been proposed to generate 3D pore space images. A series of 2D sections can be combined to form a 3D image. However, this is limited by the impossibility of preparing cross sections with a spacing of less than about 10μm (Dullien, 1992). Recently, the use of a focused ion beam technique (Tomutsa and Radmilovic, 2003) overcomes the resolution problem and it allows sub-micron image to be constructed. Non-destructive X-ray computed microtomography (Spanne et al., 1994) is another approach to image a 3D pore space directly at resolutions of around a micron. The resolution is, however, not sufficient to image the sub-micron size pores that are abundant in carbonates. The sub-micron structures of real rocks have been studied using laser scanning confocal microscopy (Fredrich, 1999). It has also limited ability to penetrate solid materials. In the absence of higher resolution 3D images, reconstructions from readily available 2D microscopic images such as scanning electron microscopy (SEM) are the only viable alternative

.

2D high-resolution images provide important geometrical properties such as the porosity and typical patterns. Based on the information extracted from 2D images, one

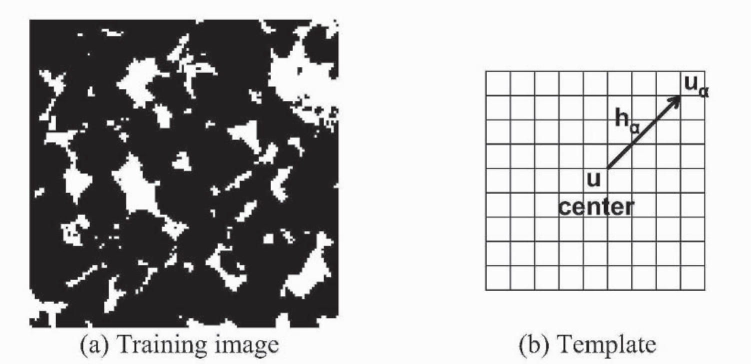
O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 763–768.
© 2005 *Springer. Printed in the Netherlands.*

promising way is to reconstruct the porous medium by modeling the geological process by which it was made (Bryant and Blunt, 1992, Bakke and Øren, 1997, Pilotti, 2000). Although this process-based reconstruction is general and possible to reproduce the long-range void connectivity, there are many systems for which the process-based reconstruction is very difficult to apply. For instance, for many carbonates it would be complex to use a process-based method that mimics the geological history involving the sedimentation of irregular shapes followed by significant compaction, dissolution and reaction (Lucia, 1999). In these cases it is necessary to find another approach to generate a pore space representation. We have reconstructed geologically realistic pore space structures using the multiple-point statistical technique (Okabe and Blunt, 2004a, 2004b), which uses higher order information (Caers, 2001, Strebelle et al., 2003). One key aspect of the work is the proper selection of the multiple-point statistics to reproduce satisfactory images. In previous work (Okabe and Blunt, 2004a, 2004b), we studied sandstones and showed that the long-range connectivity of the pore space was better reproduced. Since the method is suitable for any material, including those with sub-micron structures, we apply the method to a carbonate rock in addition to sandstones. The reconstructed 3D pore structures are tested by calculating percolation probability and predicting permeability using the lattice-Boltzmann method. Further details of the methodology can be found in Okabe and Blunt (2004b).

2 Multiple-point statistics reconstruction

Multiple-point statistics cannot be inferred from sparse data; their inference requires a densely and regularly sampled training image describing the geometries expected to exist in the real structure. Microscope images at the pore scale can be used as training images. In our application only two phases are used – void and solid phase. The method to reconstruct a 3D image from 2D information is an extended version of the multiple-point statistics approach that was developed by (Srivastava, 1992, Caers, 2001, Strebelle et al., 2003). We assume isotropy in orthogonal direction to generate a 3D image using multiple-point statistics measured on a 2D plane. Especially, we assume only a 2D plane is available to reconstruct 3D images in this study. In our rock sample, void-void autocorrelation functions (ACF) are identical for X, Y and Z directions (ACF of Z direction can be only measured for a sandstone sample using a micro-CT image); therefore, we can use this assumption. The training image and the template to capture patterns (multiple-point statistics) used is shown in Figure 1. The major extension of the method is the rotation of the measured statistics by 90 degrees, which allows us to generate a 3D structure. Since the multiple-point statistics method is well-established in geostatistics, we will not repeat the standard procedure to generate 2D images from 2D training images in detail. The procedure consists of three steps: (1) extracting multiple-point statistics from the training image; (2) probability calculation on each orthogonal plane using conditioning data; and (3) pattern reproduction using the probability weighted by the number of conditioning data on each plane.

Here we explain how to generate a 3D image using a 2D training image. This is different from the method proposed by Journel (2002). After extracting every possible pattern in the training image, every unit voxel in a 3D domain is visited once randomly. At every voxel in order to assign pore or grain phase, three principal orthogonal planes, XY, XZ and YZ intersecting the designated voxel are used to find conditioning data on these planes one by one. Consideration of the orthogonal planes is important to reproduce proper connectivity. The process, which is equivalent to the running of 2D MPS simulation for single plane, estimates each probability of the phase at the voxel on the different planes and three measured probabilities are linearly weighted by the number of conditioning data on each plane to obtain a single probability on the voxel. Finally, the phase at the voxel is assigned based on this weighted probability to generate a 3D image as assumed isotropy in orthogonal plains (Figure 2). If anisotropy is expected to exist in 3D, multi-orientation thin sections can be used as training images. There is less conditioning data during the initial stage of the reproduction. In this case, the porosity value can be used as the probability.



(a) Training image (b) Template

Figure 1. (a) An example of a training image taken from a micro-CT image of Berea sandstone with a porosity of 0.177 (128^2 pixels). The pore space is shown white and the grain black. The resolution of the image is 10μm/pixel. (b) A 9 × 9 template used to capture multiple-point statistics. The training image is scanned and each occurrence of any possible patterns of void space and solid is recorded. We also use a succession of larger templates using a form of multigrid simulation (Strebelle et al., 2003).

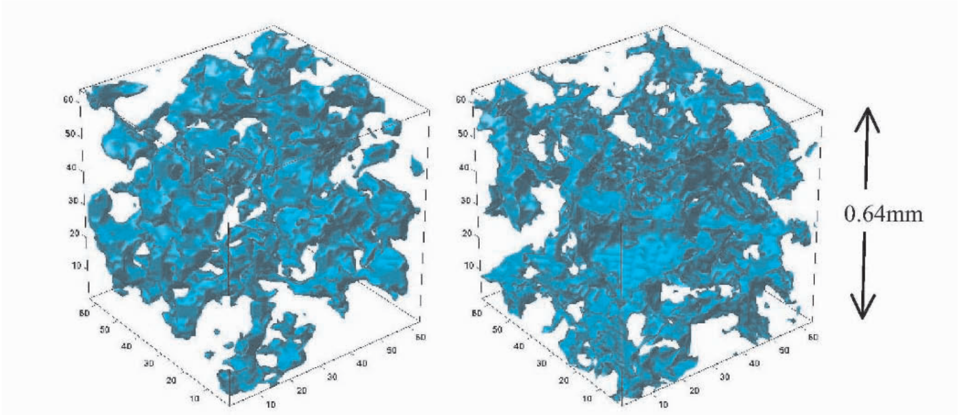


Figure 2. A subgrid of 3D image of reconstructed Berea sandstone (left, ϕ =0.1747) compared with that of the micro-CT image (right, ϕ =0.1781).

3 Percolation probability

A key aspect of our reconstruction method is the possibility to reproduce long-range connectivity. A quantitative characterization of the connectivity is provided by the local percolation probabilities or fraction of percolating cells (Hilfer, 2002) defined by

$$p_3(L) = \frac{1}{m}\sum_r \Lambda_3(r, L)$$

where *m* is the number of measurement and $\Lambda_3(r, L)$ is an indicator of percolation.

$$\Lambda_3(r, L) = \begin{cases} 1 \text{ if } M(r, L) \text{ percolates } 3 \text{ directions} \\ 0 \text{ otherwise} \end{cases}$$

A measurement cube *M(r, L)* of sidelength L centered at position *r* is used to calculate the condition of continuous connectedness from one face to opposite face by percolation theory (Stauffer and Aharony, 1994). In 3D discretized media, 26 nearest neighbors are used to measure the void connectedness. This property shows considerable difference between different reconstruction approaches. Figure 3 shows the reproduction of long-range connectivity by our method. This figure also plots the fraction of percolating cells for Berea sandstone reconstruction using simulated annealing, which matched traditional low-order properties such as porosity and two-point correlation functions, and using process-based reconstruction (Øren and Bakke, 2003). In this figure the reference measured by micro-CT and the process-based method are similar but differ from that for the structure generated using simulated annealing. This figure shows that reconstruction methods based on the low-order correlation functions fail to reproduce the long-range connectivity of porous media, while the process-based method successfully reproduces the connectivity. Our multiple-point statistics method significantly improves the connectivity over the two-point statistics method, although the pore space is still less well connected than the reference image.

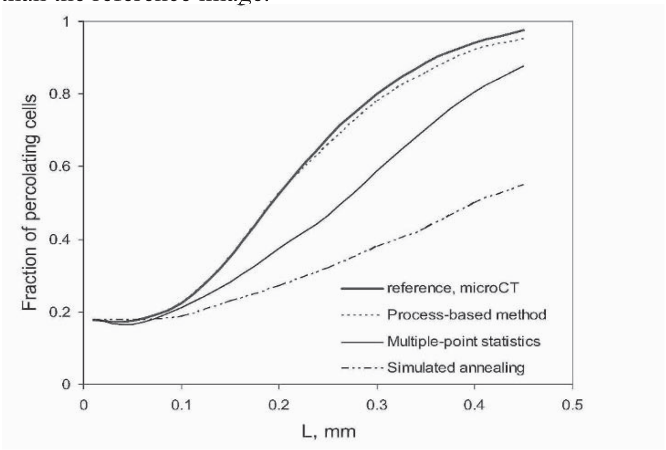


Figure 3. Fraction of percolating cells for images using different reconstruction methods. Notice that incorporating higher-order information in the reconstruction significantly improves the long-range connectivity of the pore space, although it still performs less well than process-based reconstruction methods. The data except multiple-point statistics and micro-CT are taken from (Øren and Bakke, 2003).

4 Flow properties

The lattice-Boltzmann method (LBM) provides a good approximation to solutions of the Navier-Stokes equations using a parallel and efficient algorithm that readily accommodates complex boundaries, as encountered in porous media (Buckles et al., 1994). Therefore, the LBM is used to calculate single-phase permeability to examine the reconstructed structure. This is a convenient way to assess the structures if no microtomographic image of the rock is available. The bounce-back scheme at walls is used to obtain no-slip velocity conditions and the flow field is computed using periodic boundary conditions.

The computed permeabilities of the reconstructed microstructures are listed in Table 1. Although the value for the carbonates is overestimated from the experimental permeability, the estimation is good considering the significant size difference between reconstructed images and the experimental sample. Larger training images can capture more statistics and may produce more realistic images with similar permeability values to the experiment. In addition more information, such as several thin section images and multi-orientation thin section images may improve the results.

Table 1. Computed permeabilities using the LBM.

Rock	Experiment		Computed permeability by LBM, md	
Sample	Porosity	Permeability, md	micro-CT	reconstruction
Berea	0.178	1100	1346	1274
Carbonate	0.318	6.7	N/A	19.8

5 Conclusions

A multiple-point statistics method using 2D thin sections to generate 3D pore-space representations of the rocks has been tested. The microstructures of the rocks were reconstructed and their permeabilities simulated by the lattice-Boltzmann method were compared with the experimental values. The predicted permeabilities were in good agreement with experiment data. In this study, a combination of a small 2D image and a 9×9 template with multigrid simulation was successful to capture typical patterns seen in 2D images. The reconstruction can be improved using additional information, such as higher-order information with large templates and several thin-section images including multi-orientation images if the medium is anisotropic, at the expense of more computer power and memory.

Future work will be devoted to application of the method to more rocks including carbonates, as well as the generation of topologically equivalent networks from 3D images. From the networks, predictions of capillary pressure and relative permeabilities for samples of arbitrary wettability can be made using pore-scale modeling (Blunt et al., 2002, Valvatne and Blunt, 2004).

Acknowledgements

The authors would like to express their gratitude to Japan Oil, Gas and Metals National Corporation (JOGMEC) for financial and technical support of our research and to acknowledge JOGMEC for granting permission to publish this paper with the use of the lattice-Boltzmann simulation code. We thank David Stern (ExxonMobil) and Pål-Eric Øren (Statoil) for sharing the experimental data with us, and Sebastien Strebelle for his comments. We also thank the sponsors of the Imperial College Consortium on Pore-Scale Modelling for financial support.

References

Bakke, S. and Øren, P. E., 3-D pore-scale modelling of sandstones and flow simulations in the pore networks, *SPE Journal,* **2,** 1997, p. 136-149.

Blunt, M. J., Jackson, M. D., Piri, M. and Valvatne, P. H., Detailed physics, predictive capabilities and macroscopic consequences for pore-network models of multiphase flow, *Advances in Water Resources,* **25,** 2002, p. 1069-1089.

Bryant, S. and Blunt, M., Prediction of Relative Permeability in Simple Porous-Media, *Physical Review A,* **46,** 1992, p. 2004-2011.

Buckles, J. J., Hazlett, R. D., Chen, S. Y., Eggert, K. G. and Soll, W. E., Toward Improved Prediction of Reservoir Flow Performance - simulating oil and water flows at the pore scale, *Los Alamos Science,* **22,** 1994, p. 112-120.

Caers, J., Geostatistical reservoir modelling using statistical pattern recognition, *Journal of Petroleum Science and Engineering,* **29,** 2001, p. 177-188.

Dullien, F. A. L., *Porous Media: Fluid Transport and Pore Structure,* Academic Press, San Diego. 1992.

Fredrich, J. T., 3D imaging of porous media using laser scanning confocal microscopy with application to microscale transport processes, *Physics and Chemistry of the Earth Part A-Solid Earth and Geodesy,* **24,** 1999, p. 551-561.

Hilfer, R., Review on scale dependent characterization of the microstructure of porous media, *Transport in Porous Media,* **46,** 2002, p. 373-390.

Journel, A. G., Combining knowledge from diverse sources: An alternative to traditional data independence hypotheses, *Mathematical Geology,* **34,** 2002, p. 573-596.

Lucia, F. J., *Carbonate reservoir characterization,* Springer, Berlin, Germany. 1999.

Okabe, H. and Blunt, M. J., Pore space reconstruction using multiple-point statistics, *Journal of Petroleum Science and Engineering***,** 2004a, in press.

Okabe, H. and Blunt, M. J., Prediction of permeability for porous media reconstructed using multiple-point statistics, *Physical Review E,* **70,** 2004b, in press.

Øren, P. E. and Bakke, S., Reconstruction of Berea sandstone and pore-scale modelling of wettability effects, *Journal of Petroleum Science and Engineering,* **39,** 2003, p. 177-199.

Pilotti, M., Reconstruction of clastic porous media, *Transport in Porous Media,* **41,** 2000, p. 359-364.

Spanne, P., Thovert, J. F., Jacquin, C. J., Lindquist, W. B., Jones, K. W. and Adler, P. M., Synchrotron Computed Microtomography of Porous-Media - Topology and Transports, *Physical Review Letters,* **73,** 1994, p. 2001-2004.

Srivastava, M., Iterative methods for spatial simulation, *In Report 5, Stanford Center for Reservoir Forecasting,* Stanford, CA, 1992.

Stauffer, D. and Aharony, A., *Introduction to Percolation Theory, Revised 2nd Edition,* Taylor & Francis, London & New York. 1994.

Strebelle, S., Payrazyan, K. and Caers, J., Modeling of a deepwater turbidite reservoir conditional to seismic data using principal component analysis and multiple-point geostatistics, *SPE Journal,* **8,** 2003, p. 227-235.

Tomutsa, L. and Radmilovic, V., Focussed Ion Beam Assisted Three-dimensional rock Imaging at Submicron-scale, *International Symposium of the Society of Core Analysts***,** 2003, SCA2003-47.

Valvatne, P. H. and Blunt, M. J., Predictive pore-scale modeling of two-phase flow in mixed-wet media, *Water Resources Research,* **40,** 2004, W07406, doi:10.1029/2003WR002627, 2004.

LOCAL UPDATING OF RESERVOIR PROPERTIES FOR PRODUCTION DATA INTEGRATION

LINAN ZHANG, LUCIANE B. CUNHA and CLAYTON V. DEUTSCH
Department of Civil & Environmental Engineering, 220 CEB,
University of Alberta, Canada, T6G 2G7

Abstract. A methodology is proposed that integrates historical production data into large reservoir models by the local updating of the permeability field. The focus is on conditioning a proposed initial model to injection/production rate and pressure history in an iterative fashion. Integrating flow simulation and kriging algorithms within an optimization process based on linearized formulas of reservoir behaviour with property and numerically calculated sensitivity coefficients constitutes the proposed methodology. This method makes it possible to condition the permeability distributions to injection/production rate and pressure history from large reservoirs with complex heterogeneities and changes of well system at the same time. Discussions show that sensitivity coefficients change with time/iterations and that using the linearized formula to get the optimal property changes at all master point locations is a valid strategy.

1 Introduction

There is a challenge to condition reservoir property models to production data for large scale fields with a long production/injection history accounting for realistic field conditions. Direct calculation schemes are avoided considering that they are often limited to 2-D single-phase flow. Stochastic approaches such as simulated annealing or genetic algorithms (Deutsch, 2002; Cunha, et. al., 1996) require a lot of simulation runs, making them practically unfeasible for large scale applications. Algorithms and software for production data integration based on hydrogeological developments such as sequential self-calibration (Wen, et. al., 1998; Wen, et. al., 2002) have not been proven applicable in complex reservoir settings with multiphase flow, 3-D structure and changing well conditions. Streamline-based methods have been used in large reservoirs to condition the property models to observed production rates or water cut at wells (Qassab, et. al., 2003; Agarwal and Martin, 2003; Tarun and Srinivasan, 2003), but in general, they need a finite deference method to create pressure fields so that it is difficult for these methods to condition the property models to observed well bottom-hole pressure for real large reservoirs with multiphase flow, 3-D structure and changing well conditions. The convergence of results for gradual deformation methods is slow so that lots of iterations are needed for large 3-D models (Hu, 2002; Feraille, et. al., 2003). Regularization methods like Bayesian based techniques need reliable prior information that is difficult to guarantee in many cases (Shah, et. al., 1978).

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 769–774.

There is a need for a novel computational efficient production data integration method that: (1) integrates well bottom pressure and production rate simultaneously by limited flow simulation runs, and (2) keeps a high accuracy as much as possible in large complex 3-D reservoir models with high heterogeneous property models, multiple phases, complex well system change and long history of production and injection.

2 Basic idea and general procedure of the proposed methodology

Our basic idea consists on the numerical calculation of the sensitivity coefficients on the basis of two flow simulations – an initial base case and a single sensitivity case. With this, we substitute the difficult analytical calculation of the sensitivity coefficients by a simple algorithm. The approximate sensitivity coefficients, which are used to locally update the property models, are then used to obtain optimal changes at master point locations by optimization with the linearized formulas of reservoir response changes(p-p_0, q-q_0) and reservoir property change(Δk), p-$p_0 \approx (\partial p/\partial k)\Delta k$ and q-$q_0 \approx (\partial q/\partial k)\Delta k$. Subscript "0" denotes the foundational model. The procedure is iterated until the results are satisfied or can not be improved much. The overall procedure of the proposed methodology can be summarized as follows:

At first, select an initial conditional geostatistical realization as the base model that reproduces all of the static data possible, run a flow simulation with the base model and calculate the mismatch in pressure and fractional flow rates between simulation results and historical data.

Then consider the following outer optimization loop:

- Choose one location or multiple locations to perturb based on the local mismatch at well locations – areas with greater mismatch are given a greater probability of being chosen for perturbation;
- Perturb the permeability – either by 0.5 or 1.5 perturbation factor since there is no use in making too small of a change;
- Propagate the change to all locations in the grid system, which really means the locations within the range of correlation of the changed value. The perturbation location and range may change with iteration;
- Create the perturbed model;
- Run a second flow simulation with the perturbed model and calculate the numerical sensitivity coefficients;
- Calculate optimal changes to reservoir properties at master point locations and propagate to the entire grid system;
- Run another flow simulation to establish the updated model, which may be the new base model for next iteration.
- Calculate the mismatch.

Repeat the optimization loop until the results are satisfied or can not be improved

The simulation runs involved in the methodology of production data integration proposed in this work are to be performed using the ECLIPSE flow simulator. This allows the consideration of complex geometry and heterogeneity of reservoir models as well as realistic well scheduling. However, if the finite flow simulation runs turn to be excessively costly, it is always possible to use a streamline flow simulator instead.

The formula to calculate the sensitivity coefficients of reservoir responses with respect to the permeability change are as follows:

$$SP^i_{K_h,m,w,t} = \frac{\partial p_{w,t,m}}{\partial K_h(\mathbf{u}_m)} = \frac{\widehat{p}_{w,t,m} - p^i_{w,t}}{\Delta \widehat{K}_h(\mathbf{u}_m)} = \frac{\Delta p^i_{w,t,m}}{\Delta \widehat{K}_h(\mathbf{u}_m)}$$

$$SQ^i_{K_h,m,w,t} = \frac{\partial q_{w,t,m}}{\partial K_h(\mathbf{u}_m)} = \frac{\widehat{q}_{w,t,m} - q^i_{w,t}}{\Delta \widehat{K}_h(\mathbf{u}_m)} = \frac{\Delta q^i_{w,t,m}}{\Delta \widehat{K}_h(\mathbf{u}_m)}$$

where $SP^i_{K_h,m,w,t}$ and $SQ^i_{K_h,m,w,t}$ are the sensitivity coefficients of pressure and rate at the well with index *w* and time *t* for iteration *i* with respect to horizontal permeability change $\Delta\widehat{K}_h$ at the location $\mathbf{u}_m$, respectively; $\widehat{p}_{w,t,m}$ and $\widehat{q}_{w,t,m}$ are the flow simulation results of pressure and rate at the well with index *w* and time *t* with the perturbed model by perturbing permeability only at the location $\mathbf{u}_m$, respectively; $p^i_{w,t}$ and $q^i_{w,t}$ are the simulation results of pressure and rate at the well with index *w* and time *t* with the foundation model at iteration *i*, respectively; $\Delta p^i_{w,t,m}$ and $\Delta q^i_{w,t,m}$ refer to the changes of pressure and fractional flow rate introduced by the perturbation at the location $\mathbf{u}_m$ without considering the other perturbations.

For one perturbation location at each iteration, the differences of well bottom hole pressure and fractional rate between the foundation model and the perturbed model at one iteration can be used to calculate the sensitivity coefficients directly. However, for multiple perturbation locations at each iteration, the changes of pressure and production rate at wells are the total effect caused by the joint permeability changes propagated from the multiple perturbation locations. There is a need to calculate the approximate changes of pressure and production rates caused by the permeability change propagated from one perturbation location based on the permeability values at perturbation locations and the distances between the objective well and perturbation locations.

The expectation is that after 5-20 iterations by using the proposed methodology, the number of wells with high mismatch and the highest mismatch level at wells would be reduced.

Two main features of the methodology distinguish this method from others: 1. numerically calculated sensitivity coefficients of pressure and flow rate subject to changes in porosity and permeability are used in the optimization to get the optimal property changes; 2. integrates pressure data and oil rate data to reservoir models at same time for large reservoirs with multiple phase, 3-D structure and changing well conditions by limited simulation runs.

3 Behaviour of Sensitivity Coefficients

Sensitivity coefficients of well bottom pressure and production rate subject to the property change are very important parameters in the methodology. Here the behavior of the sensitivity coefficients was studied by comparing the calculated sensitivity coefficients at Well 1 between the first two iterations in an application. Well 1 was a producer at the beginning and was converted into injector later around time of 6100. The perturbation locations, perturbation ranges and perturbation factors are the same for the two iterations. The results are shown in Figure 1. From Figure 1, we can see that the sensitivity coefficients at the well in the production period change with time and decline with iteration. For the injection period, the change of the sensitivity coefficients of well bottom pressure is more complicated. This means that we can not use one set of sensitivity coefficients for all time and all iterations.

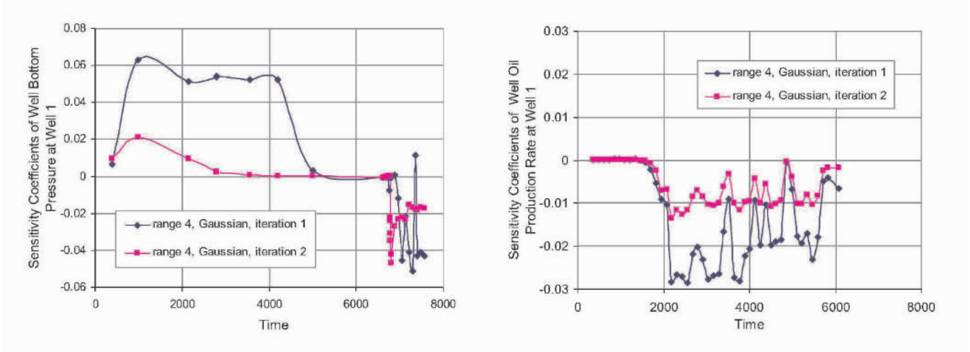


Figure 1. The behaviour of sensitivity coefficients of well bottom hole pressure and oil production rate subject to the permeability change at the grid block with Well 1 for the two iterations.

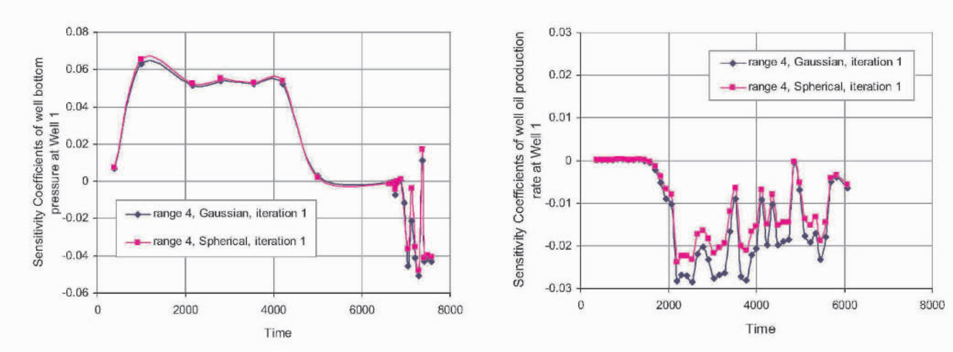
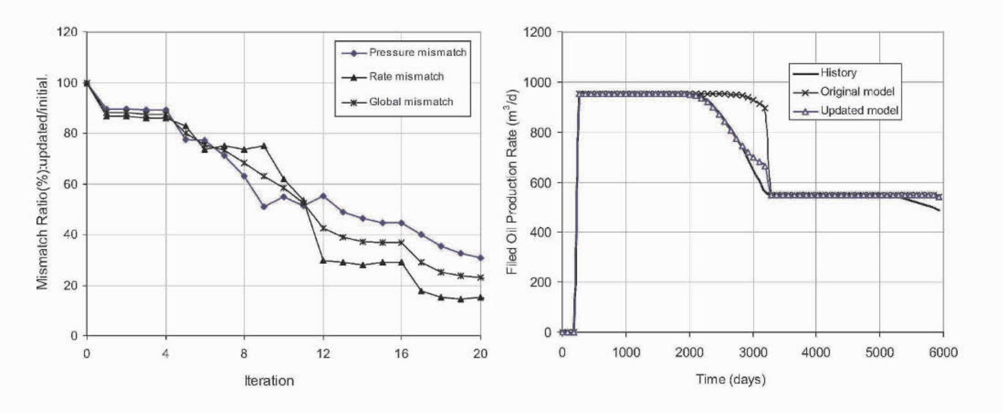


***Figure 2**.* The behaviour of sensitivity coefficients of well bottom hole pressure and oil production rate subject to permeability change at the grid block with Well 1 for the different perturbation variogram types.

From Figure 2, we can see that the perturbation variogram has a larger effect on the sensitivity coefficients of oil production rate but little effect on the sensitivity coefficients of the well bottom hole pressure. The perturbation with a variogram of Gaussian type gets larger absolute values of sensitivity coefficients of oil production rate than that with a variogram of spherical type. Considering that there is no large difference between the sensitivity coefficients of well bottom hole pressure, the perturbation with a variogram of Gaussian type may provide better results.

4. An Application of the Proposed Methodology

The proposed methodology was applied to a synthetic reservoir with 9 wells and production/injection history of 6025 days. “True” permeability and porosity models were the post-processed realizations generated from sequential Gaussian simulation/co-simulation by setting permeability and porosity as zero at the grid blocks with permeability values lower than 100md. The results from flow simulation with “true” permeability and porosity models were used as production historical data. Well liquid production rate and water injection rate were set as input parameters in flow simulation. The initial model of permeability for the methodology was generated by sequential Gaussian simulation with different random seeds from “true” models based on the well data. One perturbation location was selected at each iteration in the application. The porosity models used in flow simulation were generated by co-simulation with the correlation coefficient of 0.7 to permeability models. The results of mismatch change with iterations in the application of the methodology are shown in Figure 3(a). It can be seen that after 20 iterations, the mismatch in well bottom pressure of the updated model decreased by 69.31% from the initial model, the mismatch in oil production rate of decreased by 84.62%, the global mismatch decreased by 76.96%. Figure 3(b) shows that the updated model gets a better history match for field oil production rate. Therefore, the methodology can decrease the mismatch in well bottom hole pressure and oil production rate at the same time with a limited number of flow simulation runs.



(a) Mismatch change (b) Field oil production rate

Figure 3. Mismatch evolution with iteration number and comparative field oil production rate for a synthetic case example.

This page intentionally blank

5 Conclusions

The proposed method combines flow simulation and kriging algorithms together with an optimal technology in order to use less number of flow simulations for conditioning a proposed initial property model to fractional flow rate and pressure history at same time by an iterative scheme with numerically calculated sensitivity coefficients. The perturbation locations are selected based on the local mismatch at each well and some master point locations are used as reference positions to calculate the pressure and fractional flow rate sensitivity coefficients subject to changes in porosity and permeability. The optimal changes of porosity and permeability at the master point locations are obtained by minimizing the global mismatch related to reservoir responses of pressure and fractional flow rates calculated by linearized formulas on property change, and then are propagated to the whole grid system by kriging.

The discussion shows that we can not use one set of sensitivity coefficients of well bottom hole pressure and oil production rate subject to property change for all iterations. The application demonstrates that the methodology can reduce pressure mismatch and rate mismatch with a limited number of flow simulation runs. Additional investigation is needed in order to increase the methodology efficiency.

Acknowledgements

We would like to acknowledge the industry sponsors of the Centre for Computational Geostatistics at the University of Alberta and the National Science and Engineering Research of Canada (NSERC) for supporting this research.

References

Deutsch, C.V., *Geostatistical Reservoir Modeling*; Oxford University Press, New York, 2002.

Cunha, L.B., Oliver D.S., Redner, R.A. and Reynolds, A.C., *A Hybrid Markov Chain Monte Carlo method for generating permeability fields conditioned to multiwell pressure data and prior information*; SPE 36566 presented at the SPE Annual Technical conference and exhibition, Denver, Oct. 6-9, 1996.

Wen, X. H., Deutsch, C. V. and Cullick, A.S., *High-resolution reservoir models integrating multiplewell production data*; SPEJ , 1998, p. 344-355.

Wen X. H., Tran T. T., and Behrens R. A. and Gomez-Hernandez J. J., *Production data integration in sand/shale reservoirs using sequential self-calibration and GeoMorphing: A comparison*, SPEJ, 2002, p. 255-265.

Qassab, H., Khalifa, M., Afaleg, N. and Ali, H, *Streamline-based production data integration under realistic field conditions:Experience in a giant middle-eastern reservoir*; SPE 84079 presented at the SPE Annual Technical conference and exhibition, Denver, Oct. 5-8, 2003.

Agarwal B., Martin B., *A streamline-based method for assisted history matching applied to an Arabian Gulf Field*; SPE 84462 presented at the SPE Annual Technical conference and exhibition, Denver, Oct. 5-8, 2003.

Tarun K., Srinivasan S., *Iterative integration of dynamic data in reservoir models*; SPE 84592 presented at the SPE Annual Technical conference and exhibition, Denver, Oct. 5-8, 2003.

Hu, L.Y., *Combination of Dependent Realizations within the gradual deformation method*; Mathematical Geology, 2002, p. 953-964.

Feraille, M., Roggero, F., Manceau, E., Hu, L.Y. and Zabalza-Mezghani I., *Application of Advanced history matching techniques to an integrated field cases*; SPE 84463 presented at the SPE Annual Technical conference and exhibition, Denver, Oct. 5-8, 2003.

Shah P. C., Gavalas G. R. and Seinfeld J. H., *Error Analysis in History Matching: The Optimum Level of Parametrization*; SPE Journal, , Jun. 1978, p.219-228.

ENVIRONMENTAL

COMPARISON OF MODEL BASED GEOSTATISTICAL METHODS IN ECOLOGY : APPLICATION TO FIN WHALE SPATIAL DISTRIBUTION IN NORTHWESTERN MEDITERRANEAN SEA

PASCAL MONESTIEZ, LAURENT DUBROCA [2], EMILIE BONNIN [3],
JEAN-PIERRE DURBEC [3] and CHRISTOPHE GUINET [2]
INRA, Unité de Biométrie, Domaine Saint-Paul,
Site Agroparc, 84914 Avignon cedex 9, France.
[2] *Centre d'Etudes Biologiques de Chizé, CNRS,*
79360 Villiers en Bois, France.
[3] *Centre d'Océanologie de Marseille, Université de la Méditerranée,*
Campus de Luminy, Case 901, 13288 Marseille Cedex 9, France.

Abstract. Characterizing spatial distribution of wild species as fin whales is a major issue to protect these populations and study their interaction with their environment. Accurate maps may be difficult to obtain with very heterogeneous observation efforts and unfrequent sightings. This paper proposes to compare two geostatistical methods associated with the Poisson distribution which models the observation process. First, assuming few weak hypotheses on the distribution of abundance, we improved the experimental variogram estimate using weights that are derived from expected variances and proposed a bias correction that accounts for the Poisson observation process. The kriging system was also modified to interpolate directly the underlying abundance better than data themselves. Second the Bayesian approach proposed by Diggle in 1998 was run on the same dataset. In both case results were substantially improved compared to classical geostatistics. Advantages and drawbacks of each method are then compared and discussed.

1 Introduction

In the Mediterranean Sea, the fin whale (*Balaenoptera physalus*, Linné 1758) is the largest marine predator commonly observed (Notarbartolo di Sciara et al., 2003). Several hundred to several thousand individuals were estimated to be present in the western Mediterranean Sea during summer (Forcada *et al.*, 1996). However, common does not mean frequent. A study by Gannier (2002) gave an indication of their summer abundance in the Corso-Ligurian basin with sightings ranging from 0.6 to 9.0 whales per 100 km of transect over the period 1991-2000.

To regularize such rare sightings and pool data from different sources, counts are usually summed over small spatial cells for which cumulated observation effort

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 777–786.

is also quantified. Geostatistical modelling and mapping of this type of data address several methodological questions; how to deal with the variability from Poisson type distribution and spatial heterogeneity of observation efforts, how to handle a high proportion of zero values together with few rare high values which result from very heavy tailed distributions ?

Some of these questions have been already discussed in literature. Christensen *et al.* (2002) proposed to fit transformation from the Box-Cox family when data are positive valued and distribution skewed. If it seems adapted to rainfall data or to geochemical variables, it does not solve the problem of high proportion of zeros. Diggle *et al.* (1999) introduced what he called "model based geostatistics" which is a Generalized Linear Mixed Model (GLMM) where the random effect is a spatial Gaussian process. This framework is well adapted to inhomogeneous Poisson data distributions. However, the Bayesian framework and the computer intensive MCMC do not facilitate its use by non statistician.

Oliver *et al.* (1993) proposed a specific kriging and a bias correction for the experimental variogram when data follow a Binomial distribution. A case study in epidemiology, mapping of childhood cancer risk, is presented. The Binomial kriging takes into account the discrete nature of the data with a large proportion of zeros and spatial heterogeneity of the n - number of trials - binomial parameter.

In this paper we present a method adapted to the Poisson case which followed a similar strategy than the Oliver's one for the Binomial. This led to a modified kriging - called Poisson Ordinary Kriging in the following - and a bias corrected experimental variogram. Moreover, we introduced a weight system to improve experimental variogram estimate. Then our method is compared on the same data set to the GLMM Bayesian approach as in Moyeed and Papritz (2002).

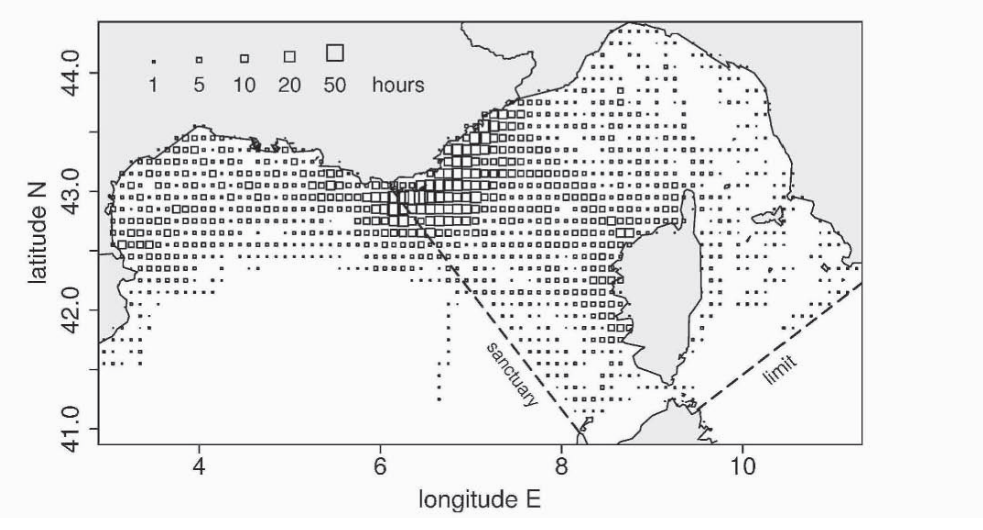


Figure 1. Map of observation data. Each cell of 0.1° × 0.1° where the sighting time was strictly positive is marked by a symbol at its center. Symbol area is proportional to the cumulated observation time.

2 Data

The fin whale sightings database used in this study merges data from different sources. Exhaustive list of contributors is given in the acknowledgement section. Only surveys for which observation effort could be quantified were used. Available sightings data covered the period 1993 to 2001. The prospected area extends from 3°E to 11.5°E and from 41°N to 44.5°N (Figure 1) including the International Cetacean Sanctuary of the Mediterranean, established on November 25th, 1999.

These surveys were conducted either along random linear transects or onboard ferries along their regular lines between France and Corsica. The number of whales reported for a given sighting was often unreliable and only sightings number was considered in this study. A GPS (Global Positioning System) recorded the vessels tracks. The July and August data for all years were then aggregated on cells of 0.1° of longitude by 0.1° of latitude ($\sim$ 90 km^2) in a regular grid. We computed in each cell the cumulated number of whale sightings, the observation effort which was defined as the total time (in hours) spent observing and the averaged number of sightings per hour of observation (Figures 1 and 2).

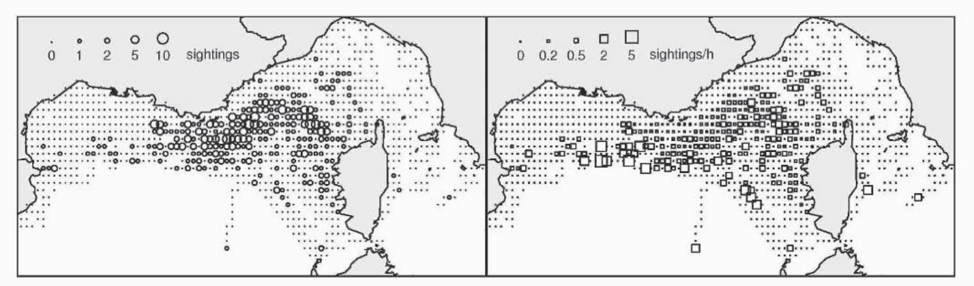


Figure 2. Maps of sightings data (left) and of averaged sightings per hour (right). Symbol areas are proportional to variables.

Histograms in Figure 3 show the skewed distributions of the three previous variables for the 1113 cells where observation times was strictly positive. The raw number of sightings ranged from 0 (864 cells) to 11, and the average number of

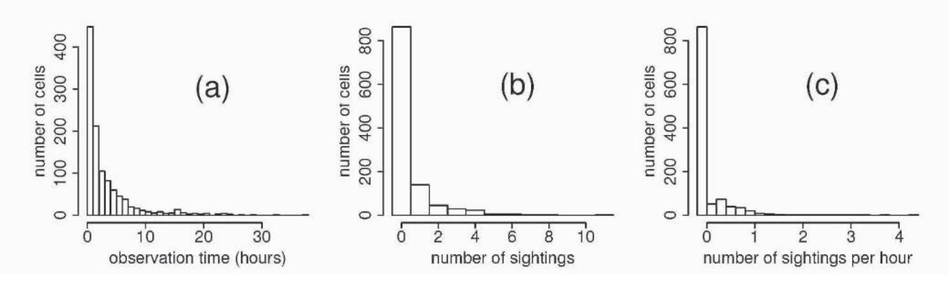


Figure 3. Histograms of observation times for cells with $t_s > 0$ (a), of raw sightings data (b) and of number of sightings per hour averaged on cells (c).

sightings per hour ranged from 0 to 4.22 with a mean estimated to 0.14 sighting/h.

The total number of sightings was 490 for a total time of 3484 hours of observation. The number of cells where fin whales were seen was 249, so effective data made of unfrequent sightings were sparsely distributed in space and locally mixed with zero values. Moreover because of great heterogeneity in observation times, the map of raw sightings data reveals as well the fin whale presence as the observation effort intensity.

3 Models and methods

For all site s belonging to domain $\mathcal{D}$, we define the random field $Z(s)$ by

$$Z(s)|Y(s) \sim \mathcal{P}\big(t(s)\; Y(s)\big) \tag{1}$$

where $Z(s)|Y(s)$ is Poisson distributed with a parameter that is the product of $t(s)$ by $Y(s)$, where $t(s)$ is the observation time (in hours) at site s, $Y(s)$ is proportional to a relative animal abundance at site s and measures the expectation of sightings for a unit observation time.

$Y(s)$ is a positive random field honoring order two stationarity, with mean m, variance σ_Y^2 and covariance function $C_Y(s-s')$ or variogram $\gamma_Y(s-s')$. Conditionally on Y, the random variables $Z(s)$ are mutually independent.

To simplify notations, $Z(s)$, $Y(s)$ and $t(s)$ will be noted in the following Z_s, Y_s and t_s respectively. In kriging systems, $C_{ss'}$ denotes the covariance $C_Y(s-s')$.

3.1 MODEL I : POISSON ORDINARY KRIGING

In this section Y_s is distribution free with unknown mean and we have only to assume that $Y_s \geq 0$.

3.1.1 *Expectation and variance of Z_s*

As $Z(s)|Y_s \sim \mathcal{P}\big(t_s Y_s\big)$ it follows directly that :

$$\begin{aligned}
\mathsf{E}[Z_s|Y_s] &= t_s Y_s & \mathsf{E}[Z_s] &= m t_s \\
\mathsf{Var}[Z_s|Y_s] &= t_s Y_s & & \\
\mathsf{E}\big[(Z_s)^2\big|Y_s\big] &= t_s Y_s + \big(t_s Y_s\big)^2 & \mathsf{Var}[Z_s] &= t_s^2 \sigma_Y^2 + m t_s \\
\mathsf{E}\Big[Z_s\, Z_{s'}\Big|Y\Big] &= \delta_{ss'}\, t_s Y_s + t_s t_{s'} Y_s Y_{s'} & &
\end{aligned} \tag{2}$$

where $\delta_{ss'}$ is the Kronecker delta which is 1 if $s=s'$ and 0 otherwise.

3.1.2 *Expectation and variance of $\left(\frac{Z_s}{t_s} - \frac{Z_{s'}}{t_{s'}}\right)$*

In order to characterize the relationship between the variograms of Z and Y, we develop the expressions of the two first moments of $\big(\frac{Z_s}{t_s} - \frac{Z_{s'}}{t_{s'}}\big)$. After checking that the expectation is null, it results :

$$\frac{1}{2}\,\mathsf{E}\left[\left(\frac{Z_s}{t_s} - \frac{Z_{s'}}{t_{s'}}\right)^2\right] = \frac{m}{2}\left(\frac{t_s + t_{s'}}{t_s\, t_{s'}}\right) - \delta_{ss'}\,\frac{m}{t_s} + \gamma_Y(s-s') \tag{3}$$

Let $\gamma_Z(s-s')$ denote the non-stationary theoretical variogram corresponding to the random field $\big(Z_s/t_s\big)$, we get for $s \neq s'$ the relationship :

$$\gamma_Y(s-s') = \gamma_Z(s-s') - \frac{m}{2}\left(\frac{t_s + t_{s'}}{t_s\, t_{s'}}\right) \tag{4}$$

We can check for $s = s'$ that equation (4) reduces to $\gamma_Y(0) = \gamma_Z(0) = 0$

Furthermore the conditional variance is given by :

$$\mathsf{E}\left[\mathsf{Var}\left[\frac{Z_s}{t_s} - \frac{Z_{s'}}{t_{s'}}\middle| Y\right]\right] = \mathsf{E}\left[\frac{Y_s}{t_s} + \frac{Y_{s'}}{t_{s'}}\right] = m\left(\frac{t_s + t_{s'}}{t_s\, t_{s'}}\right) \tag{5}$$

3.1.3 *Estimation of* $\gamma_Y(h)$

Let Z_α,$\alpha = 1, \dots, n$ be the n measurements of $Z(s_\alpha)$ obtained during observation times t_α. The expression of a modified experimental variogram can be derived from (4) and (5).

$$\gamma^*_Y(h) = \frac{1}{N(h)} \sum_{\alpha=1}^{n}\sum_{\beta=1}^{n} \frac{t_\alpha\, t_\beta}{t_\alpha + t_\beta}\left[\frac{1}{2}\left(\frac{Z_\alpha}{t_\alpha} - \frac{Z_\beta}{t_\beta}\right)^2 - \frac{m^*}{2}\left(\frac{t_\alpha + t_\beta}{t_\alpha\, t_\beta}\right)\right]\mathbb{1}_{d_{\alpha\beta}\sim h} \tag{6}$$

where $\mathbb{1}_{d_{\alpha\beta}\sim h}$ is the indicator function of pairs (s_α, s_β) whose distance is close to h, where $N(h) = \sum_{\alpha,\beta} \frac{t_\alpha\, t_\beta}{t_\alpha + t_\beta}\, \mathbb{1}_{d_{\alpha\beta}\sim h}$ is a normalizing constant and where m^* is an estimate of the mean of Y.

The bias correction term $-\frac{m^*}{2}\left(\frac{t_s + t_{s'}}{t_s\, t_{s'}}\right)$ derives directly from (4).

The weights $\frac{t_s\, t_{s'}}{t_s + t_{s'}}$ are introduced to homogenize the variance of differences terms $\left(\frac{Z_s}{t_s} - \frac{Z_{s'}}{t_{s'}}\right)$ by dividing them by a weight proportional to their standard deviation $\sqrt{m\frac{t_s + t_{s'}}{t_s\, t_{s'}}}$ given by (5). When simplifying (6) we get for $h \neq 0$:

$$\gamma^*_Y(h) = \frac{1}{2\,N(h)} \sum_{\alpha,\beta}\left(\frac{t_\alpha\, t_\beta}{t_\alpha + t_\beta}\left(\frac{Z_\alpha}{t_\alpha} - \frac{Z_\beta}{t_\beta}\right)^2 - m^*\right)\mathbb{1}_{d_{\alpha\beta}\sim h} \tag{7}$$

3.1.4 *Ordinary kriging of* Y_o

Poisson Ordinary Kriging at any site s_o is a linear predictor of Y_o combining the observed data Z_α weighted by observation times t_α. The mean of Y_o is supposed unknown.

$$Y^*_o = \sum_{\alpha=1}^{n} \lambda_\alpha \frac{Z_\alpha}{t_\alpha} \tag{8}$$

Unbiasedness constrain λ_α to sum up to one as for classical Ordinary Kriging (OK). The variance of the error of prediction, i.e. the MSEP if unbiased, was obtained by first expressing $\mathsf{E}[(Y^*_o - Y_o)^2|Y]$ and then deconditioning :

$$\mathsf{E}\big[(Y^*_o - Y_o)^2\big] = \sigma^2_Y + \sum_{\alpha=1}^{n} \frac{\lambda^2_\alpha}{t_\alpha} m + \sum_{\alpha=1}^{n}\sum_{\beta=1}^{n} \lambda_\alpha \lambda_\beta\, C_{\alpha\beta} - 2\sum_{\alpha=1}^{n} \lambda_\alpha C_{\alpha o} \tag{9}$$

By minimizing this expression (9) on λ_i's with the unbiasedness constraint, we obtain the following kriging system of $(n+1)$ equations where μ is the Lagrange multiplier.

$$\left\{\begin{array}{l} \displaystyle\sum_{\beta=1}^{n} \lambda_\beta C_{\alpha\beta} + \lambda_\alpha \frac{m}{t_\alpha} + \mu = C_{\alpha o} \quad \text{pour} \quad \alpha = 1, \ldots, n \\ \displaystyle\sum_{\alpha=1}^{n} \lambda_\alpha = 1 \end{array}\right. \tag{10}$$

The expression of the kriging variance resulting from system (10) reduced after calculation to the same expression than for classical OK. However the kriging variance map may be very different due to changes in resulting λ_α. Calculations and intermediate results of section 3.1 are detailed in Monestiez *et al.* (2004).

3.2 MODEL II : SPATIAL GLMM

We added an hypothesis on the random field $Y(s)$ that becomes lognormal:

$$\begin{aligned} Z(s)|Y(s) &\sim \mathcal{P}\big(t(s)\, Y(s)\big) \\ \log\big(Y(s)\big) &= \beta + S(s) \end{aligned} \tag{11}$$

where, following Diggle's notations, $S(s)$ is a zero-mean Gaussian random field with variance σ_ϵ^2, covariance function $\sigma_\epsilon^2\, \rho(s-s')$ and where $\rho(s-s')$ is a parametric autocorrelation function with scale parameter φ.

3.2.1 *Spatial GLMM and Bayesian framework*

The model can be interpreted as a spatial generalized mixed model (GLMM), where β is a fixed effect reduced to a simple mean effect, and S a random effect whose parameters are $\theta = (\sigma_\epsilon^2, \varphi)$. The link function is here the log transform.

Let S_α denote $S(s_\alpha)$ at a data site s_α ($\alpha \in 1, \ldots, n$), $S_{-\alpha}$ the vector of S_1 to S_n with element S_α removed, Z the vector of observation data Z_α and S_o the value of S at any point s_o where a prediction is wanted. In such context, the kriging predictor Y_o^* may be replaced by $\hat{Y}_o = \exp(\hat{\beta} + \widehat{S_o})$ where $\widehat{S_o}$ would be ideally $\mathsf{E}\big[S_o|Z\big]$. Since the number of data is large and S spatially dependent, it is clear that special methods will be needed to solve this problem.

Diggle *et al.* (1998) proposed a Bayesian framework coupled with MCMC methods (Robert and Casella, 1999). MCMC is a natural tool since the conditional distribution of Z given S and the marginal distribution of S derive directly from the model (11) and the conditional independence of Z given Y.

3.2.2 *Posterior simulations and predictions*

To implement our MCMC scheme we need to generate random samples from the posterior distributions $\pi(\theta|S, Z, \beta)$, $\pi(\beta|S, Z, \theta)$ and $\pi(S_s|S_{-s}, Z, \beta, \theta)$ for inference, and from $\pi(S(s_o)|S, Z, \beta, \theta)$ for prediction in s_o. In this paper, we do not have the possibility to expose the whole method, so we refer to the original paper of Diggle *et al.* (1998, pp 306–309) or to Christensen and Waagepetersen (2002, pp 282–283) who detailed the expressions of conditional densities and the different steps of the MCMC scheme using a Metropolis-Hastings algorithm. An application on count data can be found in Wikle (2002) who addresses the problem of mapping bird breeding over the continental United States.

The computation was performed with the software R (R Development Core Team, 2004) and the package GeoRglm (Christensen and Ribeiro, 2002). After some tests run on simulated examples, we finally ran one million of iterations, storing parameters and simulated prediction grids only every 1000 iterations. The priors for β and σ_ϵ^2 were non-informative uniforms while the scale parameter φ was fixed using the same covariance model shape than for kriging (stable model with power parameter set to 1.5). Cpu time reached several hours on recent office PC running on Windows XP, a time which remains acceptable but was more than 500 times longer than the mapping by Poisson Ordinary Kriging on R.

4 Results

4.1 EXPERIMENTAL VARIOGRAMS AND VARIOGRAM FITTING

We computed the standard experimental variogram on Z_s/t_s and the one defined by equation (7) on Y_s. Figure 4 shows clearly the effect of both corrections, pair weighting and bias. Weights on pairs led to a more regular experimental variogram with a lower sill. After bias correction we can assume that γ_Y has no more nugget effect. A stable variogram model (equation 12) was thus fitted :

$$\gamma_Y(h) = c\Big(1 - \exp\big(-(h/a)^d\big)\Big) \tag{12}$$

with parameters $c = 0.043$, $a = 28.4$ and $d = 1.51$; intermediate model in smoothness between an exponential variogram ($d = 1$) and a gaussian variogram ($d = 2$).

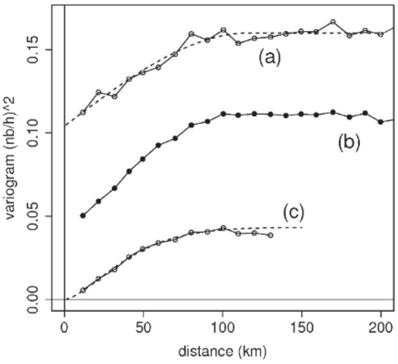


Figure 4. Experimental variograms on sightings per hour Z_α/t_α. (a) standard one and fitted spherical model; (b) experimental variogram when introducing the weights on pairs; (c) experimental variogram from Eqn. (7) and fitted stable model.

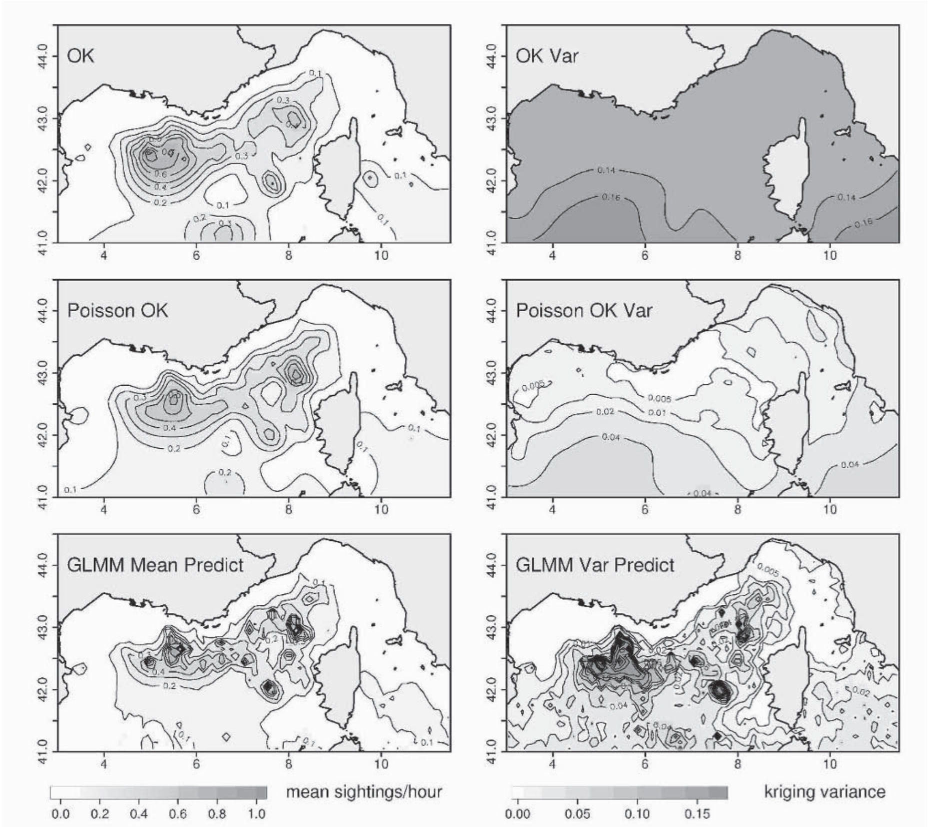


Figure 5. Maps of predictions of Y^* (left) and of prediction variances (right) of fin whale sightings per hour; by Ordinary Kriging of Z_α/t_α (top), Poisson Ordinary Kriging (middle) and GLMM simulations (bottom).

4.2 KRIGING AND PREDICTIONS

We defined a grid for prediction of Y values that reproduced the sample grid. The prediction grid with elementary cell of 0.1° by 0.1° extends also from 3°E to 11.5°E and from 41°N to 44.5°N. All points beyond the coastline were removed so it remained 2020 points to predict. Among these points, 1113 were located in a cell with observed Z and t. Kriging system was established using unique neighborhood.

Figure 5 displays maps of predictions, and of variances of predictions, for OK on Z_α/t_α - with the variogram model of Figure 4a - for Poisson OK and for GLMM.

Global patterns of abundance are similar, however maps obtained by the two model based methods show more local spatial patterns and differs only for higher values with a smoothing for the kriging. Conversely, the three methods gave completely different variance maps. The first map from kriging is flat in the central region where data and prediction grid overlap. The Poisson OK variances are substantially smaller and modulated by the observation times. The GLMM variance of prediction map differs because of the lognormal hypothesis. Variances varies with the predicted values, with highest values that are ten times the kriging ones for large $\widehat{Y}$ (ranging from 0.0001 to 0.47). In GLMM variance also accounted for observation times but this effect is masked by the previous one.

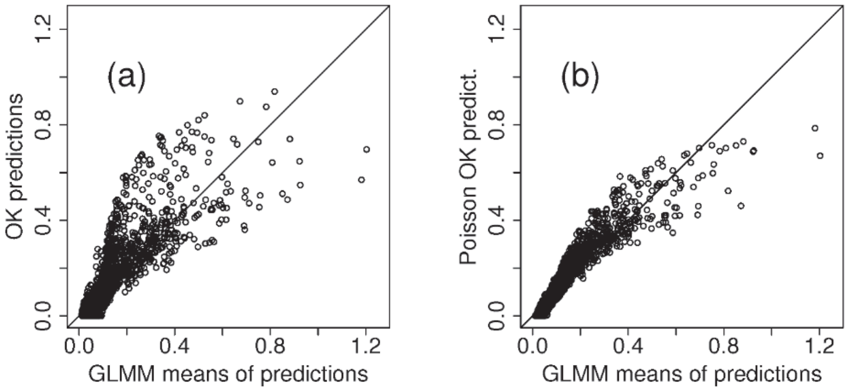


Figure 6. Plots of predictions of whale sightings per hour by Ordinary Kriging versus GLMM (left), and by Poisson Ordinary Kriging versus GLMM (right).

Cross validation is not available since true values of Y are unknown even in presence of data, so we compared predictions in Figure 6. We can consider that GLMM and Poisson OK gave equivalent results for values lower than 0.4, i.e. for more than 90% of predictions. The curvature of the cloud in Figure 6b is probably due to exponential transform of S used in the GLMM prediction of Y.

5 Discussion and conclusions

The mapping method we proposed which is a specific kriging written for Poisson distribution case do not raise more difficulties than Ordinary Kriging. However, modification of standard software is necessary, which is easy when using open statistical software as R (R Development Core Team, 2004). Specific R functions were written by the authors. But this can become a problem with some programs plugged in GIS commercial software.

The Poisson Ordinary Kriging gave maps that were adapted to ecologists needs and consistent with others studies. The fin whale data set was quite extreme considering the heterogeneity of observation times and the very low values for the sighting frequency so we believe that the proposed method will be able to give satisfactory results in other ecological surveys. It is a real advantage to remain as simple as OK and to not have to introduce distributional hypothesis on animal abundance, sharing the robustness of OK.

A drawback of our model-free spatial abundance is the possibility of negative mapped values. For fin whales it happened exceptionally in region of lowest density with negative predictions whose absolute values were negligible. A simple way to solve the difficulty was to set at zero the rare negative predictions, so the MSEP was globally reduced. However, if such negative predictions based on positive data becomes more frequent, this could suggest that the chosen variogram model is wrong, especially for short distances.

Differences between Poisson OK and Diggle or Wilke's GLMM come from the lognormal hypothesis which are probably not relevant here for higher levels of concentration of fin whales. If it is probably correct to model a proportionality between variance and mean for $Z|Y$ because of the Poisson observation process, there is no reason to expect such similar relation on Y as strictly modelled by lognormal distribution.

Acknowledgements

The authors thank all the people and institutions who contributed to the sightings data collection: the CRMM, the CETUS, the CIESM, the CIRCÉ, the Delphinia Sea Conservation, the EPHE, the GECEM, the IFREMER, the Musée Océanographique de Monaco, the RIMMO, the SACLANTCEN, the SNCM and the Swiss Cetacean Society.

References

Christensen, O.F., Diggle, P.J., Ribeiro Jr., P.J., 2002. Analysing Positive-Valued Spatial Data: The Transformed Gaussian Model. In: Monestiez, P. *et al.* (Eds), geoENV III - Geostatistics for Environmental Applications. Kluwer Academic Publishers, Dordrecht, pp. 287–298.

Christensen, O. F. and Ribeiro Jr., P. J., 2002, geoRglm: A package for generalised linear spatial models, R-NEWS 2, 26–28. http://cran.R-project.org/doc/Rnews

Christensen, O. F., Waagepetersen, R., 2002. Bayesian Prediction of Spatial Count Data Using Generalized Linear Mixed Models. Biometrics 58, 280–286.

Diggle, P.J., Tawn, J.A., Moyeed, R.A., 1998. Model Based Geostatistics. Appl. Statist. 47, 299–350.

Forcada, J., Agiolar, A., Hammond, P., Pastor, X., Aguila, R., 1996. Distribution and Abundance of Fin Whales (*Balaenoptera physalus*) in the Western Mediterranean Sea during the Summer. Journal of Zoology 238, 23–24.

Gannier, A., 2002. Summer Distribution of Fin Whales (*Balenoptera physalus*) in the Northwestern Mediterranean Marine Mammals Sanctuary. Revue d'Ecologie (Terre Vie) 57, 135–150.

Monestiez, P., Dubroca, L., Bonnin, E., Guinet, C., Durbec, J.-P., 2004. Geostatistical Modelling of Spatial Distribution of *Balenoptera physalus* in the Northwestern Mediterranean Sea from Sparse Count Data and Heterogeneous Observation Efforts. Technical Report, INRA, Avignon.

Moyeed, R.A., Papritz, A., 2002. An Empirical Comparison of Kriging Methods for Nonlinear Spatial Point Prediction. Mathematical Geology 34, 365–386.

Notarbartolo di Sciara, G., Zanardelli, M., Jahoda, M., Panigada, S., Airoldi, S., 2003. The Fin Whale *Balaenoptera physalus* (L. 1758) in the Mediterranean Sea. Mammal Rev. 33, 105–150.

Oliver, M. A., Lajaunie, C., Webster, R., Muir, K. R., Mann, J. R., 1993. Estimating the Risk of Chilhood Cancer. In: Saores, A. (Ed.), Geostatistics Troia '92. Kluwer Academic Publishers, Dordrecht, pp. 899–910.

R Development Core Team, 2004. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, iSBN 3-900051-00-3. http://www.R-project.org

Robert, C., Casella, G., 1999. Monte Carlo Statistical Methods. Springer-Verlag, New York.

Wikle, C. K., 2002. Spatial Modeling of Count Data: A Case Study in Modelling Breeding Bird Survey Data on Large Spatial Domains. In: Lawson, A., Denison, D. (Eds.), Spatial Cluster Modelling. Chapman and Hall, London, pp. 199–209.

SIMULATION-BASED ASSESSMENT OF A GEOSTATISTICAL APPROACH FOR ESTIMATION AND MAPPING OF THE RISK OF CANCER

PIERRE GOOVAERTS
BioMedware, Inc. 516 North State Street, Ann Arbor, MI 48104

Abstract. This paper presents a geostatistical methodology that accounts for spatially varying population sizes and spatial patterns in the processing of cancer mortality data. The binomial cokriging approach is adapted to the situation where the variance of observed rates is smaller than expected under the binomial model, thereby avoiding negative estimates for the semivariogram of the risk. Simulation studies are conducted using lung cancer mortality rates measured over two different geographies: New England counties and US State Economic Areas. For both datasets and different spatial patterns for the risk (i.e. random, spatially structured with and without nugget effect) the proposed approach generally leads to more accurate risk estimates than traditional binomial cokriging, empirical Bayes smothers or local means.

1 Introduction

Cancer mortality maps are important tools in health research, allowing the identification of spatial patterns and clusters that often stimulate research to elucidate causative relationships (Jacquez, 1998; Goovaerts, 2005). The analysis is however frequently hampered by the presence of noise in mortality data, which is often caused by unreliable extreme relative risks estimated over small areas, such as census tracts (Mungiole *et al.,* 1999). Statistical smoothing algorithms have been developed to filter local small-scale variations (i.e. changes occurring over short distances) from mortality maps, enhancing larger-scale regional trends (Talbot *et al.,* 2000). A limitation of the smoothers reported in today's health science literature is that they cannot be tailored easily to the pattern of variability displayed by the data. For example, inverse distance methods ignore important features such as anisotropy or range of spatial correlation.

Geostatistics (Goovaerts, 1997) provides a set of statistical tools for analyzing and mapping data distributed in space and time. There have however been relatively few applications of geostatistics to cancer data, with alternative solutions to the problem of non-stationarity of the variance caused by spatially varying population sizes. In his book (p.385-402), Cressie (1993) proposed a two-step transform of the data to remove first the mean-variance dependence of the data and next the heteroscedasticity. Traditional variography was then applied to the transformed residuals. In their study on the risk of childhood cancer in the West Midlands of England, Oliver *et al.* (1998) developed an approach that accounted for spatial heterogeneity in the population of children to estimate the semivariogram of the "risk of developing cancer" from the semivariogram

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 787–796.

of observed mortality rates. Binomial cokriging was then used to produce a map of cancer risk. In their review paper Gotway and Young (2002) showed how block kriging can account for differing supports in spatial prediction (aggregation and disaggregation approach), allowing the analysis of relationships between disease and pollution data recorded over different geographies. More recently, Goovaerts *et al.* (2005) presented an adaptation of semivariogram and factorial kriging analysis that accounts for spatially varying population size in the processing of cancer mortality data.

Capitalizing on previous results on binomial cokriging and weighted semivariograms of cancer mortality data, this paper presents a geostatistical filtering approach for estimating cancer risk from observed rates. The risk is here defined as the probability of a person contracting the disease within a specified period (Waller and Gotway, 2004). Unlike most of the earlier work published in the geostatistical literature, prediction performances of the proposed filtering technique is assessed using simulation studies. Lung cancer mortality data recorded over New England counties (1950-1994) are analyzed geostatistically, and simulated rates are generated under a Binomial distribution model. The study is then extended to the 506 US State Economic Areas.

2 Geostatistical Analysis of Cancer Rates

For a given number *N* of entities (e.g. counties, states, electoral ward), denote the number of recorded mortality cases by $d(\mathbf{u}_\alpha)$ and the size of the population at risk $n(\mathbf{u}_\alpha)$. Following most authors (Cressie, 1993; Oliver *et al*., 1998; Christakos and Lai, 1997), entities are referenced geographically by their centroids (or seats) with the vector of spatial coordinates $\mathbf{u}_\alpha=(x_\alpha,y_\alpha)$, which means that the actual spatial support (i.e. size and shape of the county or ward) is ignored in the analysis. The empirical or observed mortality rates are then denoted as $z(\mathbf{u}_\alpha)=d(\mathbf{u}_\alpha)/n(\mathbf{u}_\alpha)$. Figure 1 shows an example for 295 counties of 12 New England States. The directly age-adjusted mortality rates for lung cancer, as provided by the new Atlas of United States mortality (Pickle *et al.,* 1999), are displayed for white males (1950-1994 period). The scattergram shows how the size of the population at risk varies among counties (from 2,185 to 716,000) and the greater variability of rates recorded for small population sizes.

The rates recorded at N=295 counties can be modeled as the sum of the risk of developing cancer and a random component (error term ε) due to spatially varying population size, $n(\mathbf{u}_\alpha)$:

$$Z(\mathbf{u}_\alpha)=R(\mathbf{u}_\alpha)+\varepsilon(\mathbf{u}_\alpha) \qquad \alpha=1,\ldots,N \qquad (1)$$

Conditionally to a fixed risk function, the counts $d(\mathbf{u}_\alpha)$ follow then a binomial distribution with parameters $R(\mathbf{u}_\alpha)$ and $n(\mathbf{u}_\alpha)$. In other words, there are two possible outcomes: having cancer or not, with $R(\mathbf{u}_\alpha)$ being the probability of having the disease. The following relations are satisfied:

$$E[\varepsilon(\mathbf{u}_\alpha)]=0 \quad \text{and} \quad Var[\varepsilon(\mathbf{u}_\alpha)]=R(\mathbf{u}_\alpha)\times\{1-R(\mathbf{u}_\alpha)\}/n(\mathbf{u}_\alpha) \qquad (2)$$

$$E[Z(\mathbf{u}_\alpha)]= E[R(\mathbf{u}_\alpha)]=\mu \quad \text{and} \quad Var[Z(\mathbf{u}_\alpha)]=Var[R(\mathbf{u}_\alpha)]+Var[\varepsilon(\mathbf{u}_\alpha)] \qquad (3)$$

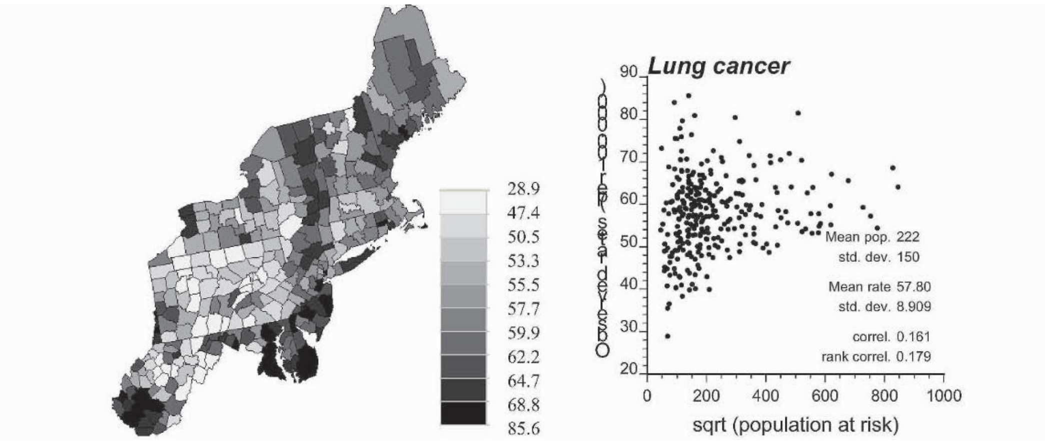


Figure 1. Map of lung cancer mortality rates recorded over the period 1950-1994, and their relationship to the size of the population at risk (white males).

For estimation purpose and in agreement with Oliver *et al.* (1998), the variance of the error term can be approximated as $\mathrm{Var}[\varepsilon(\mathbf{u}_\alpha)]=\sigma_\varepsilon^2=\mu\times(1-\mu)/n(\mathbf{u}_\alpha)$, where the mean parameter μ is estimated by the population-weighted average of rates, $\overline{z}$. The risk over a given entity with centroid $\mathbf{u}_\alpha$ is estimated from $s(\mathbf{u}_\alpha)$ neighboring observed rates as:

$$\hat{R}(\mathbf{u}_\alpha) = \sum_{i=1}^{s(\mathbf{u}_\alpha)} \lambda_i(\mathbf{u}_\alpha) z(\mathbf{u}_\alpha) \tag{4}$$

The kriging weights are solution of the following system:

$$\begin{aligned} &\sum_{j=1}^{s(\mathbf{u}_\alpha)} \lambda_j(\mathbf{u}_\alpha) C(\mathbf{u}_i\text{-}\mathbf{u}_j) + \mu(\mathbf{u}_\alpha) = C_R(\mathbf{u}_i - \mathbf{u}_\alpha) \qquad i=1,\ldots,s(\mathbf{u}_\alpha) \\ &\qquad\qquad \sum_{j=1}^{s(\mathbf{u}_\alpha)} \lambda_j(\mathbf{u}_\alpha) = 1 \end{aligned} \tag{5}$$

where $C(\mathbf{u}_i\text{-}\mathbf{u}_j)=\{1\text{-}1/n(\mathbf{u}_i)\}C_R(0)+\overline{z}\times(1\text{-}\overline{z})/n(\mathbf{u}_i)$ if $\mathbf{u}_i=\mathbf{u}_j$ and $C_R(\mathbf{u}_i\text{-}\mathbf{u}_j)$ otherwise. The addition of an "error variance" term for a zero distance accounts for variability arising from population size, leading to smaller weights for less reliable data (i.e. measured over smaller population). Note that kriging is here used to filter the noise from the observed rates aggregated to the county level, not to estimate the risk within the county itself (disaggregation procedure). There is thus no change of support and the underlying hypothesis is that all counties have the same spatial support.

System (5) requires knowledge of the covariance of the unknown risk, $C_R(\mathbf{h})$. Following the approach derived by Oliver *et al.* (1998), the unknown semivariogram of the risk and the experimental semivariogram of observed rates are related as:

$$\hat{\gamma}_R(\mathbf{h}) = \hat{\gamma}_Z(\mathbf{h}) - \tfrac{1}{2}\left\{\overline{z}(1-\overline{z}) - \hat{\sigma}_R^2\right\}\left\{\frac{1}{N(\mathbf{h})}\sum_{\alpha=1}^{N(\mathbf{h})} \frac{n(\mathbf{u}_\alpha)+n(\mathbf{u}_\alpha+\mathbf{h})}{n(\mathbf{u}_\alpha)\times n(\mathbf{u}_\alpha+\mathbf{h})}\right\} \tag{6}$$

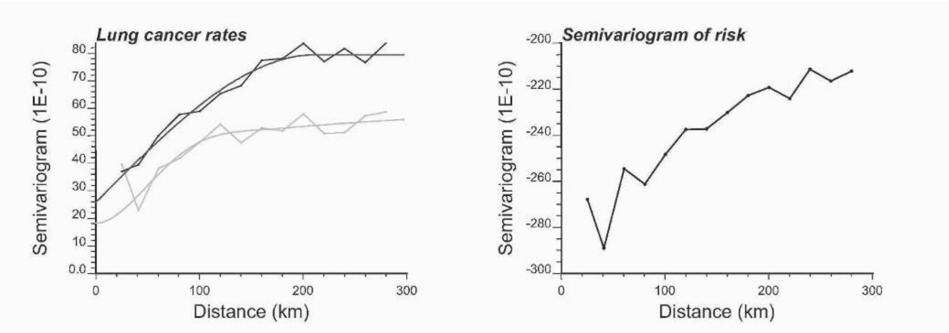


Figure 2. Semivariograms of lung cancer mortality rates with the model fitted by weighted least-squares (gray curve = population weighted semivariogram). Right graph shows the semivariogram of risk estimated according to expression (6).

An iterative procedure is used to estimate the variance of the risk $\hat{\sigma}_R^2$ which is a priori unknown, see Oliver *et al.* (1998) for a more detailed description. Application of formula (6) to New England data leads to negative values for the experimental semivariogram of the risk (see Figure 2, right graph), a disconcerting feature that has been observed on various datasets with different geographies and population sizes. According to simulation studies this problem is caused by the overestimation of the variance of the error term by the expression $\bar{z}\times(1-\bar{z})/\text{n}(\mathbf{u}_\alpha)$. In other words, all developments (1) through (6) are based on the modeling of the error term as a Binomial random variable, an assumption which may not always be consistent with the observed variability. The following empirical modification of the binomial cokriging approach is proposed to allow the use of the filtering technique in all situations, hence its implementation in a user-friendly software.

First, by analogy with Rivoirard *et al.* (2000) I propose to estimate the semivariogram of the risk by the following population-weighted semivariogram of observed rates:

$$\hat{\gamma}(\mathbf{h}) = \frac{1}{2\sum_{\alpha=1}^{N(\mathbf{h})} n(\mathbf{u}_\alpha)n(\mathbf{u}_\alpha+\mathbf{h})} \sum_{\alpha=1}^{N(\mathbf{h})} n(\mathbf{u}_\alpha)n(\mathbf{u}_\alpha+\mathbf{h})\left[z(\mathbf{u}_\alpha)-z(\mathbf{u}_\alpha+\mathbf{h})\right]^2 \qquad (7)$$

The weighting scheme attenuates the impact of data pairs that involve at least one rate computed from small population sizes, revealing structures that might be blurred by the random variability of extreme values. The weighting also tends to lower the sill of the semivariogram as well as the nugget variance, see Figure 2 (left graph).

The second modification relates to the kriging system (5) itself. In particular the term $\bar{z}\times(1-\bar{z})/\text{n}(\mathbf{u}_\alpha)$ can become disproportionately large relatively to the variance of the risk $C_R(0)$, leading to very large diagonal elements in the kriging matrix (and indirectly very large nugget effect). Such a severe understatement of the spatial correlation between rates typically results in over-smoothing since the risk becomes a simple population-weighted average of observed rates. The map of filtered rates in Figure 3 (left top graph) indeed appears much more homogeneous or smoother than the map of raw rates in Figure 1. An easy way to check for any discrepancy is to compare the sill of

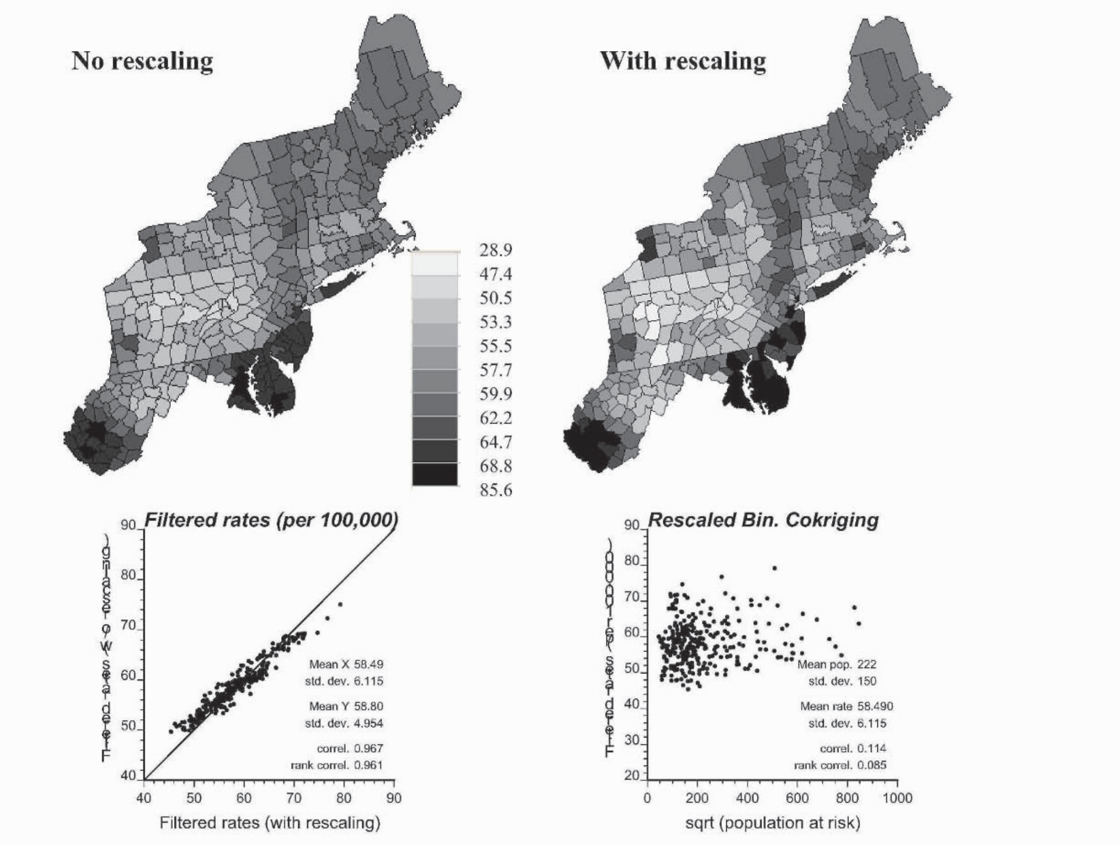


Figure 3. Maps of filtered lung cancer rates obtained by binomial cokriging with and without rescaling of kriging diagonal terms. The rescaling reduces the over-smoothing of filtered rates while still attenuating the variability among rates for small populations.

the semivariogram of observed rates $C_Z(0)$ with the value of the error variance averaged over all locations:

$$R_B = C_Z(0)\,/G \quad \text{with} \quad G = \frac{1}{N}\sum_{\alpha=1}^{N}\frac{\bar{z}(1-\bar{z})}{n(\mathbf{u}_\alpha)} \tag{8}$$

For New England data $C_Z(0)=7.937\ 10^{-9}$, while G is one order of magnitude larger $G=3.282\ 10^{-8}$, yielding a ratio $R_B=0.242$.

The proposed modification of the binomial cokriging system consists of rescaling the correction of the diagonal term to account for any discrepancy between estimates of the rate and error variances, that is $C(\mathbf{u}_i\text{-}\mathbf{u}_j)=\{1\text{-}1/n(\mathbf{u}_i)\}C_R(0)+\{\,\bar{z}\times(1\text{-}\,\bar{z}\,)/n(\mathbf{u}_i)\}R_B$. Figure 3 (right top graph) shows that this rescaling reduces the smoothing of filtered rates: the standard deviation of the distribution of filtered rates is $6.115\ 10^{-5}$ when the rescaling is performed, compared to $4.954\ 10^{-5}$ in the traditional implementation of binomial cokriging. Comparison of scattergrams of Figures 1 and 3 also indicates that extreme rates disappear after the filtering by the modified binomial cokriging algorithm.

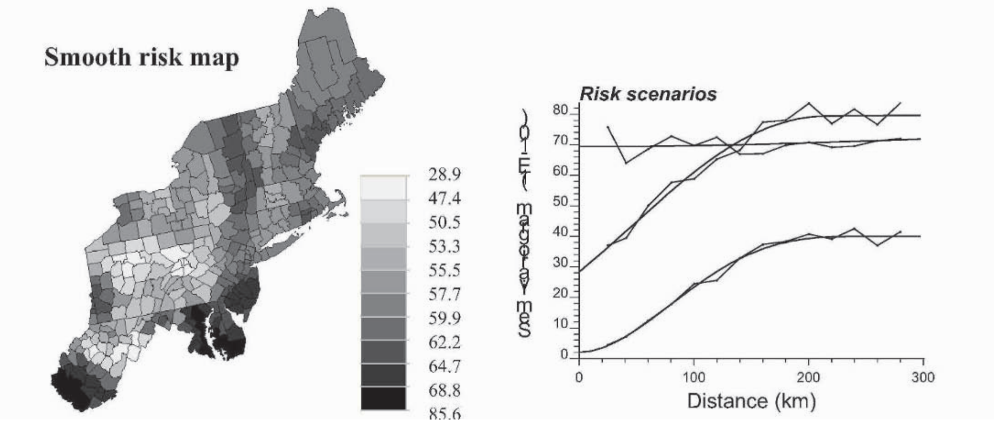


Figure 4. Smooth model for the spatial distribution of the risk of developing lung cancer obtained by a local average of the map of Figure 1, and the semivariograms of the three risk maps used in the simulation studies.

3 Simulation Study

A series of simulated maps of cancer mortality rates $\{z^{(l)}(\mathbf{u}_\alpha)\ \alpha=1,\ldots,N\}$ were generated in order to investigate the prediction performance of the use of population-weighted semivariogram and empirical rescaling of diagonal terms of the cokriging system. Three different underlying maps of risk $\{r(\mathbf{u}_\alpha)\ \alpha=1,\ldots,N\}$ were considered:

1) map of observed rates displayed in Figure 1,
2) smooth map of rates obtained by a moving population-weighted average of 6 closest neighboring counties (see Figure 4),
3) non-structured map of rates created by random shuffling of the observed rates.

The corresponding semivariograms are displayed in Figure 4. For each map of risk 100 realizations of the number of cases were generated for each county with centroid $\boldsymbol{u}_\alpha$ by random drawing of a binomial distribution with parameters $r(\mathbf{u}_\alpha)$ and $n_S(\mathbf{u}_\alpha)$. This approach is thus different from a traditional p-field simulation where the random numbers would be spatially correlated (Goovaerts, 1997). To cover a wide range of values for the rescaling factor R_B three different scenarios were considered for the population sizes used in the simulation, $n_S(\mathbf{u}_\alpha)$, and in the geostatistical analysis, $n_A(\mathbf{u}_\alpha)$: 1) $n_S(\mathbf{u}_\alpha)=n_A(\mathbf{u}_\alpha)=10\times$observed population sizes $n(\mathbf{u}_\alpha)$, 2) $n_S(\mathbf{u}_\alpha)=n_A(\mathbf{u}_\alpha)=n(\mathbf{u}_\alpha)$, and 3) $n_S(\mathbf{u}_\alpha)=10\times(n_A(\mathbf{u}_\alpha)=n(\mathbf{u}_\alpha))$. Scenario I corresponds to the simplest task in that the variance of binomial distributions is small following the arbitrary multiplication of the population size by 10; leading to a semivariogram of simulated rates close to the underlying $\gamma_R(\mathbf{h})$; see Figure 5 (left graph). Random fluctuations are more important in scenario II, leading to higher sills and less spatial structure for the semivariogram of simulated rates (Figure 5, middle graph). Unlike in scenarios I and II, the assumption made in the analysis of the third set of simulated rates is inconsistent with the actual simulation procedure; that is the variability assumed under the binomial model using the population size $n(\mathbf{u}_\alpha)$ is larger than the underlying model (since the simulation used 10 times larger population sizes). This third scenario will lead to R_B values smaller than 1

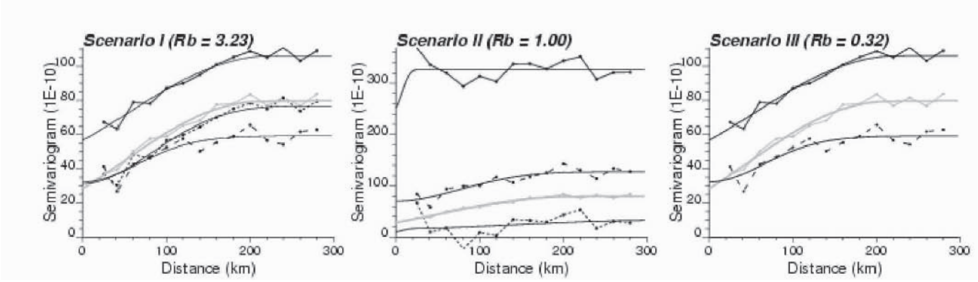


Figure 5. Semivariograms of the underlying risk (gray solid line) and rates simulated using the risk map of type 1 (i.e. observed rates) and three scenarios for the population sizes. The semivariograms are: unweighted (—), population-weighted (—), and risk semivariogram estimated according to expression (6) (…).

and negative estimates of the semivariogram of the risk, mimicking the situation observed for New England data, see Figure 5 (right graph).

Figure 5 shows the results of the variography for realization #5 generated using the risk map of type 1 the semivariogram of which is depicted by the thick gray line. Regardless of the scenario used for the population sizes, the larger sill is observed for the unweighted semivariogram of simulated rates. Incorporation of population sizes through estimator (7) reduces the sill value which can be either smaller or larger than the target risk semivariogram. The performance of the risk semivariogram estimator (6) deteriorates as the rescaling factor R_B becomes smaller, that is as the population size decreases and the Binomial model overestimates the variability that is actually observed. For scenario III, the correction applied to the semivariogram of rates is so exaggerated that the semivariogram of risk estimates are negative for all lags.

For each of the nine combinations of 3 risk maps and 3 population size scenarios, the simulated rates were filtered using binomial cokriging with and without rescaling, and three types of semivariogram estimators: underlying $\gamma_R(h)$, Oliver *et al.*'s estimator (6), and population-weighted estimator (7). Whenever the estimator (6) yielded negative values, the filtered rates were identified to a local population-weighted average of the 32 closest mortality rates, denoted $m(\mathbf{u}_\alpha)$. This latter estimator is considered as the reference filter (left column in Table 1), while binomial cokriging with the true underlying semivariogram of risk represents the best case scenario which is never encountered in practice. To include traditional non-geostatistical filters in our comparison study, empirical Bayes smoothing was also implemented. Following Waller and Gotway (2004), the global Bayes smoother of the rate at $\mathbf{u}_\alpha$ is as follows:

$$\hat{r}(\mathbf{u}_\alpha) = \lambda(\mathbf{u}_\alpha) z(\mathbf{u}_\alpha) + \left[1 - \lambda(\mathbf{u}_\alpha)\right]\bar{z} \tag{9}$$

The Bayes shrinkage factor $\lambda(\mathbf{u}_\alpha)$ is computed as:

$$\lambda(\mathbf{u}_\alpha) = \begin{cases} \dfrac{s^2 - \bar{z}/\bar{n}}{s^2 - \bar{z}/\bar{n} + \bar{z}/n(\mathbf{u}_\alpha)} & \text{if} \quad s^2 \geq \bar{z}/\bar{n} \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

where $\bar{z}$ and s^2 are the population-weighted sample mean and variance of rates, and $\bar{n}$ is the average population size across the study area. Whenever the rate $z(\mathbf{u}_\alpha)$ is based on small population sizes $n(\mathbf{u}_\alpha)$ relatively to the average $\bar{n}$, the factor $\lambda(\mathbf{u}_\alpha)$ is small and the Bayes estimate (9) is close to the global mean $\bar{z}$. In other words, the relative weight assigned to the observed rate is small since it is deemed less reliable. Local Bayes smoothers are computed similarly except that the global statistics $\bar{z}$ and s^2 are replaced by local ones computed within local search windows (i.e. estimated from the closest 32 rates in this paper). Prediction performances were quantified by the average squared differences between the filtered rates and the map of risk values used in the simulation. Results in Table 1 indicate that:

- Binomial cokriging yields more accurate estimates than a simple population-weighted local mean for all 9 scenarios, and it outperforms empirical Bayes smoothers in all cases.
- Rescaling yields less accurate risk estimates when the factor R_B is much larger than 1 (Population size I) since then the diagonal elements are inflated, leading to more smoothing effect.
- Except for the rarely encountered case of spatially random risk, the use of population-weighted semivariograms outperforms the risk semivariogram estimator of type (6). The latter yields systematically negative estimates for the Population size scenario III, which according to the value of the rescaling factor R_B is the most consistent with the observed variability.

Table 1. Prediction errors obtained on average over 100 simulations generated under three different population size scenarios and 3 types of risk map (I=observed, II=smooth, III=random). Binomial cokriging has been conducted with (w) and without (w/o) rescaling of diagonal elements of the kriging matrix. Bold numbers refer to best performances outside the ideal case where the true semivariogram of risk is known.

	Local mean	Empirical Bayes		Binomial cokriging with $\gamma_R(\mathbf{h})$						R_B
				True		Estimator (6)		Estimator (7)		
		Glob	Loc	w/o	w	w/o	w	w/o	w	
Risk I										
Size I	61.1	25.2	21.4	17.2	23.8	18.8	23.5	**17.7**	25.0	3.3
Size II	64.9	63.7	54.4	46.7	46.7	52.0	51.2	48.4	**47.9**	1.1
Size III	61.1	82.4	55.7	34.0	23.8	61.1	61.1	35.7	**25.0**	0.3
Risk II										
Size I	19.1	17.7	10.9	6.07	6.56	11.1	10.1	**6.04**	6.43	2.2
Size II	23.0	34.6	22.9	17.0	17.0	24.4	23.9	**22.3**	22.4	1.0
Size III	19.1	39.6	19.0	9.61	6.56	19.1	19.1	9.60	**6.43**	0.2
Risk III										
Size I	72.9	19.6	19.4	18.6	25.6	**18.7**	24.4	19.2	26.5	3.4
Size II	76.7	56.2	58.8	54.4	54.7	57.8	**56.1**	58.5	57.1	1.2
Size III	72.9	82.0	64.1	39.7	25.6	72.9	72.9	39.9	**26.5**	0.3

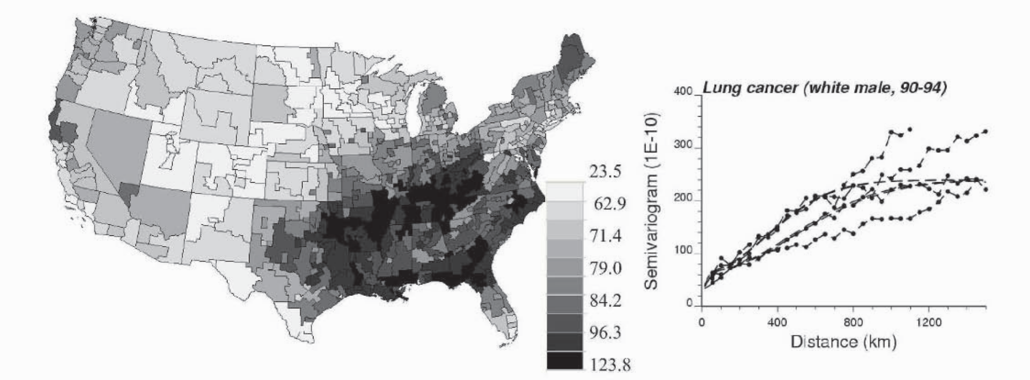


Figure 6. Map of lung cancer mortality rates recorded over the period 1990-1994 for SEA units, with the corresponding directional semivariograms.

Table 2. Prediction errors obtained on average over 100 simulations generated under population size scenario II, $n_S(\mathbf{u}_\alpha)=n_A(\mathbf{u}_\alpha)=n(\mathbf{u}_\alpha)$, and for 3 types of risk map (I=observed, II=smooth, III=random); see Table 1 for further explanations.

Risk type	Local mean	Empirical Bayes		Binomial cokriging with $\gamma_R(\mathbf{h})$						R_B
				True		Estimator (6)		Estimator (7)		
		Glob.	Loc.	w/o	w	w/o	w	w/o	w	
I	93.2	86.1	54.1	44.1	48.9	59.9	51.5	49.5	**45.6**	2.7
II	36.8	78.9	28.3	14.5	15.1	41.0	30.4	28.3	**20.1**	2.1
III	214	73.6	**73.8**	70.6	88.6	74.0	81.3	73.9	79.5	3.3

A similar simulation study was conducted over all 506 US State Economic Areas (SEA), in order to investigate the performances of the proposed approach over an area of larger extent and using rates recorded over bigger and more populated geographical units. Lung cancer rates recorded for white males during the period 1990-1994 were analyzed and they are mapped in Figure 6. The variability is clearly anisotropic with more spatial continuity along the NE-SW direction. The observed rescaling ratio is now larger than 1, R_B=1.95, which is caused by a combination of larger spatial variability (higher sill $C_Z(0)$) when the entire US is studied and larger population sizes which reduces the value of parameter G in expression (8). Nevertheless, results in Table 2 clearly demonstrate the benefit of using population weighted semivariograms and rescaled cokriging system when the risk is spatially correlated (Risk scenarios I and II).

4 Conclusions

Cancer mortality maps are used by public health officials to identify areas of excess and to guide surveillance and control activities. Quality of decision-making thus relies on an accurate quantification of risks from observed rates which can be very unreliable when computed from sparsely populated geographical units. This paper improves earlier implementation of binomial cokriging to develop an approach that is more flexible and robust with respect to misspecification of the underlying hypothesis. Simulation studies conducted under different spatial patterns of risk and population size scenarios demonstrate that the combined use of population-weighted semivariogram and rescaled cokriging system leads to more accurate estimates of the underlying risk. The implementation of the developed methodology was facilitated by the initial assumption that all geographical units are the same size, which allowed the use of geographical centroids in semivariogram estimation and kriging. This assumption is unsatisfactory when working with vastly different entities, such as SEA units over the US. A proper account of the spatial support would also allow the mapping of the risk within each unit. In addition, counts are aggregated over a given temporal period: the longer this period, the larger the smoothing of the variability in space and the greater the discrepancy between the sizes of the current population and the population that was actually exposed over this period. Underestimation of the exposed population could be the culprit for negative semivariogram estimates and this critical issue needs to be further explored.

Acknowledgements

This research was funded by grant 1R43CA105819-01 from the National Cancer Institute and benefited from discussions with Ms. Ariane Saint-Pierre during her internship with BioMedware, Inc. The views stated in this publication are those of the author and do not necessarily represent the official views of the NCI.

References

Christakos, G., and Lai, J., A study of the breast cancer dynamics in North Carolina, *Social Science & Medicine*, vol. 45(10), 1997, p. 1503-1517.

Cressie, N., *Statistics for Spatial Data*, Wiley, 1993.

Goovaerts, P., *Geostatistics for Natural Resources Evaluation*, Oxford University Press, 1997.

Goovaerts, P., Jacquez, G.M. and Greiling, D.A., Exploring scale-dependent correlations between cancer mortality rates using factorial kriging and population-weighted semivariograms, *Geographical Analysis*, 2005, in press.

Goovaerts, P., Detection of spatial clusters and outliers in cancer rates using geostatistical filters and spatial neutral models. In *geoENV V: Geostatistics for Environmental Applications*, edited by A. Soares, J. Gomez-Hernandez, and R. Froidevaux, Dordrecht: Kluwers Academic, 2005, in press.

Gotway C.A. and Young, L.J., Combining incompatible spatial data, *Journal of the American Statistical Association*, vol. 97, 2002, p. 632-648.

Jacquez, G., GIS as an enabling technology. In *GIS and Health*, edited by A. Gatrell, and M. Loytonen, London: Taylor and Francis, 1998, p. 17-28.

Mungiole, M., Pickle, L.W. and Hansen Simonson, K., Application of a weighted head-banging algorithm to mortality data maps, *Statistics in Medicine*, vol. 18, 1999, p. 3201-3209.

Oliver, M.A., Webster, R., Lajaunie, C., Muir, K.R., Parkes, S.E., Cameron, A.H., Stevens, M.C.G. and Mann, J.R., Binomial cokriging for estimating and mapping the risk of childhood cancer, *IMA Journal of Mathematics Applied in Medicine and Biology*, vol. 15, 1998, p. 279-297.

Pickle, L.W., Mungiole, M., Jones, G.K., and White, A.A., Exploring spatial patterns of mortality: the new Atlas of United States mortality, *Statistics in Medicine*, vol. 18, 1999, p. 3211-3220.

Rivoirard, J., Simmonds, J., Foote, K.G., Fernandez, P. and Bez, N., *Geostatistics for Estimating Fish Abundance*, Blackwell Science, Oxford, 2000.

Talbot, T.O., Kulldorff, M., Forand, S.P. and Haley, V.B., Evaluation of spatial filters to create smoothed maps of health data, *Statistics in Medicine*, vol. 19, 2000, p. 2399-2408.

Waller L.A., and Gotway C.A., *Applied Spatial Statistics for Public Health Data*, John Wiley and Sons, New Jersey, 2004.

AIR QUALITY ASSESSMENT USING STOCHASTIC SIMULATION AND NEURAL NETWORKS

ANA RUSSO [(1)], CARLA NUNES [(1,2)], ANA BIO [(1)], Mª JOÃO PEREIRA [(1)] AND AMILCAR SOARES [(1)]
[(1)] *Environmental Group of the Centre for Modelling Petroleum Reservoirs, CMRP/IST, Av. Rovisco Pais, 1049-001 Lisbon, Portugal. arusso@ist.utl.pt*
[(2)] *Universidade de Évora, Portugal.*

Abstract.

Since the 60's, there has been a strong industrial development in the Sines area, on the southern Atlantic coast of Portugal, including the construction of petrochemical and energy-related industries. These industries are, nowadays, responsible for substantial emissions of SO_2, NO_x, particles, VOCs and part of the ozone polluting the atmosphere. The major industries are spatially concentrated in a restricted area, very close to populated areas and natural resources. Their emissions are very similar, making the identification of individual pollutant sources and of their contributions to air pollution difficult.

In this study, the regional spatial dispersion of sulphur dioxide (SO_2) is characterized, through the combined use of diffusive tubes (Radiello Passive Samplers) and classical monitoring stations' air quality data. The objective of this study is to create a regional predictive model of the contribution of different emission sources to the pollutant concentrations captured at each monitoring station.

A two-step methodology was used in this study. First, the time series of each data pair – industrial emission and monitoring station records – was screened, in order to obtain contiguous time periods with a high contribution of that specific industrial emission to the equivalent monitoring-station measurements. For this purpose, an iterative optimisation process was developed, using a variogram between industrial emissions and monitoring-station time series as the objective function. Afterwards, probability neural networks (PNN) were applied to achieve an automatic classification of the time series into two classes: a class of (emission/monitoring station) pairs of highly correlated points and a class of pairs of points without correlation

In a second step, the relationship between time series of emissions and air quality (AQ) monitoring station records – time model – is validated for the entire area for a given period of time, using for that purpose the diffusive samplers measurements. A spatial stochastic simulation is applied to generate a set of equi-probable images of the pollutant, which relationship with the different emissions is calculated using the PNN predictive model.

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 797–807.

1 Introduction

It is well known that air pollutants at ground level can be harmful to human health, if their concentrations exceed certain limits. As pollutants accumulate in, or near, large metropolitan areas, populations are typically more exposed to unhealthy pollutant concentrations (Russo et al., 2004; Cobourn et al., 2000; Kolehmainen et al., 2000). A study that allows the identification of regional pollutant emission-receptions patterns and the quantification of the contribution of local industrial units is of great interest for the health system and environmental policy making (Russo et al., 2004; Cobourn et al., 2000; Kolehmainen et al., 2000).

Predictive modelling of the different emissions' contribution to the pollutant concentrations captured at a certain monitoring station, will allow an analysis of the impact caused in the monitoring station's area and its translation into an air quality index (Russo et al., 2004). In order to develop robust predictive air quality (AQ) models, wide-range monitoring systems are necessary. Modelling therefore often needs to be used in conjunction with other objective assessment techniques, including monitoring, emission measurement and inventories, interpolation and mapping (WHO, 1999).

Air quality monitoring can essentially be accomplished by the use of continuous automatic sensors, passive samplers, active samplers and remote sensors. The use of passive samplers offers a simple and cost-effective method of screening air quality in an area (Cruz and Campos, 2002; Mukerjee et al., 2004). The low unit costs permit sampling at numerous points in the area of interest, what is useful in highlighting "hot-spots" of high pollutant concentrations. Combined with automatic sensors, that can provide high-resolution measurements (typically hourly averages) at few points for most of the criteria pollutants (SO_2, NO_2, O_3), a spatial-temporal monitoring net may be accomplished.

2 Objectives

Briefly, the purpose of this study is to analyse possible relations between sulphur dioxide (SO_2) emissions, generated by three industrial complexes (Borealis, Petrogal and CPPE) located in the Sines area (Portugal), and AQ data colleted by three air quality monitoring stations (Sonega, Monte Chãos, Monte Velho) and also by Radiello diffusive tubes covering the Sines area, with analysis done by means of linear and non-linear modelling, as described in section 4. A predictive model of the contribution of different emission sources to the pollutant concentrations captured at each monitoring station was created using spatial information (captured by passive monitors (diffusive tubes)) combined with temporal information (captured by monitoring stations) for the same pollutants.

A two-step methodology was used in this study: i) First, the time series of each data pair – industrial emission and monitoring station records – was screened, in order to obtain contiguous time periods with a high contribution of that specific industrial emission to the equivalent monitoring-station measurements. For this purpose, an iterative optimisation process was developed, using the variogram between industrial emissions and monitoring-station time series as the objective function. Afterwards, probability neural networks (PNN's) were applied to predict the probability of pollutant emissions causing the pollutant concentrations measured at the monitoring stations; ii) In a second step, the relationship between emissions and AQ monitoring station records – time model – is validated for the entire area for a given period of time, using the diffusive samplers measurements. A spatial stochastic simulation (direct sequential simulation) is applied to generate a set of equi-probable images of the pollutant and the relationship of different emissions with local simulated values is evaluated for the entire area.

3 Study Area and Data

The main objective of this study consists in developing and implementing a methodology that allows classifying the contribution of different emission sources to air quality (AQ) in the region of Sines, Portugal (Fig. 1). Automatic sensors and passive samplers (Radiello diffusive tubes) were used in order to collected AQ data in the Sines area.

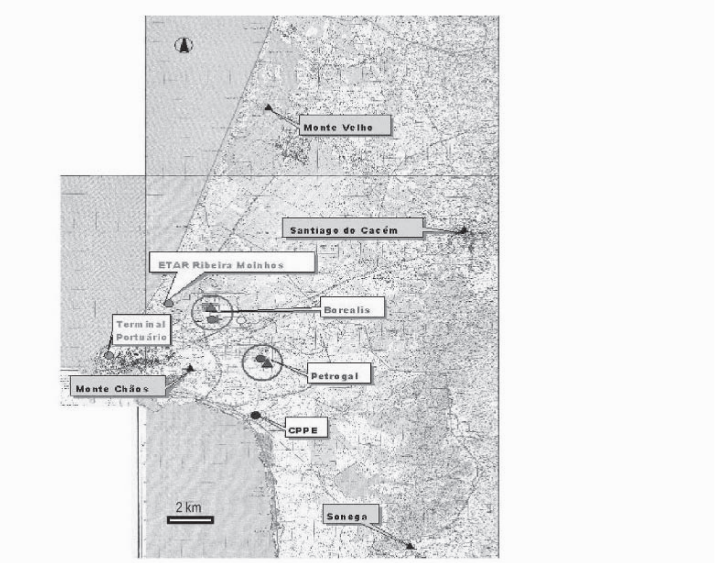


Figure 1. An overview of the Sines Peninsula (Petrogal, Borealis and CPPE industrial complexes in light gray; AQ monitoring stations in Sonega, Monte Chãos, Monte Velho in dark gray).

The case study covers an area with very different land uses: industrial, urban, rural and leisure. Although the urban area is very small, compared with the rural area, the industrial sources are of great importance and make an important contribution to long-term or peak concentrations of critical pollutants (SO_2, NO_2, O_3, etc.). Mixed occupation areas with great industrial influence should be continuously monitored and controlled, in order to prevent air quality crises.

The sulphur dioxide (SO_2) concentrations were measured in three industrial sites (Borealis, Petrogal and CPPE) and by three monitoring stations (Sonega, Monte Chãos and Monte Velho). For the purpose of this study, those measures were converted into daily averages for a period of 12 months (from 1/1/2002 to 31/12/2002) (Figs. 2 and 3). Meteorological data – wind speed and direction on an hourly basis, for the same period – were also collected and analysed (Fig. 4).

Diffusive tubes measurements of SO_2 were available for a period of 11 consecutive days (from 31/3 to 10/4/2003) (Fig. 5). The sampling period was characterized by dominant winds from north/northwest with average speeds of 10-17 km/h. The humidity levels varied between 80% and 100%. The air temperature had a typical spring variation, with an average temperature of about 15 ºC.

The available data was standardized in order to minimize the effect of different local means and variances in the evaluation of the emissions/AQ measurements relationships. Afterwards, those days, which did not have any register of data in at least one of the emission-reception stations, were deleted from the file.

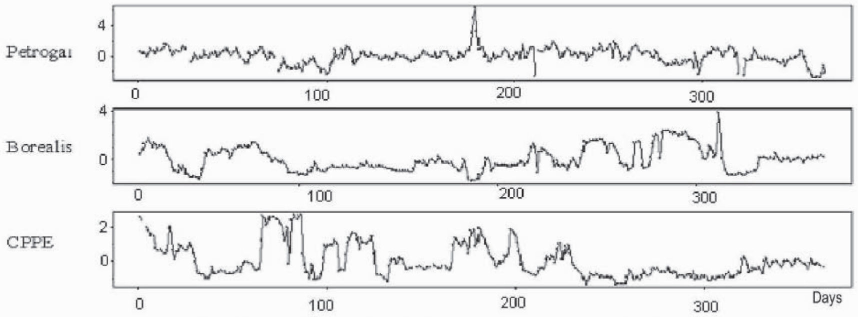


Figure 2. SO_2 (mg m^{-3}) emitted by the three industrial complexes.

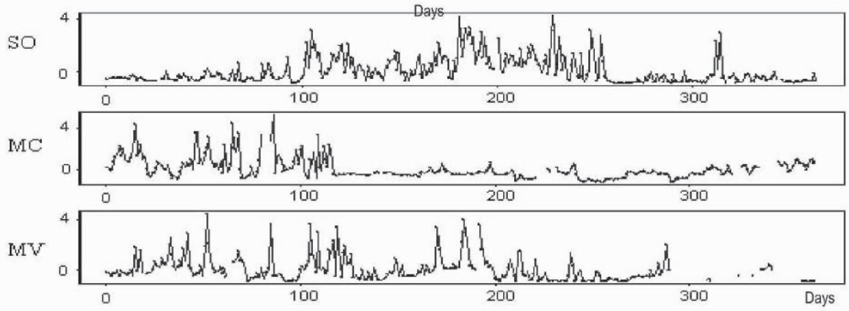


Figure 3. SO_2 (μg m^{-3}) measured by the monitoring stations (SO - Sonega, MC - Monte Chãos, MV - Monte Velho).

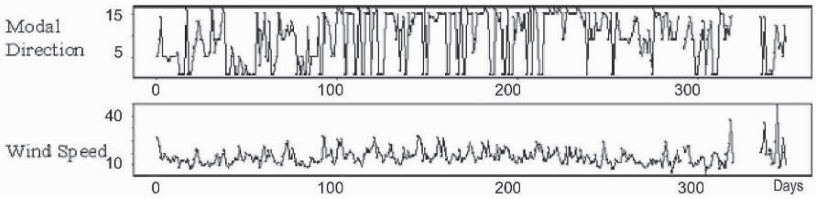


Figure 4. Wind speed (m s^{-1}) and modal wind direction registered.

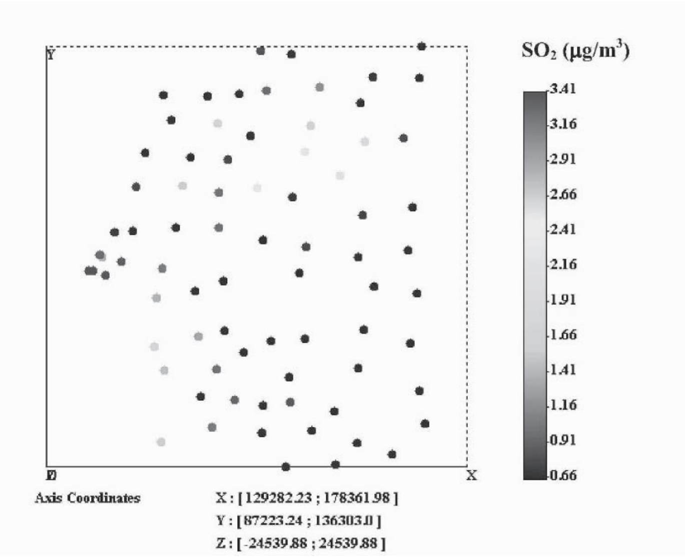

Y
SO$_2$ (µg/m^3)
3.41
3.16
2.91
2.66
2.41
2.16
1.91
1.66
1.41
1.16
0.91
0.66
D
X
Axis Coordinates
X : [129282.23 ; 178361.98]
Y : [87223.24 ; 136303.11]
Z : [-24539.88 ; 24539.88]

***Figure 5*.** Spatial SO$_2$ (µg m^{-3})dispersion measured by diffusive tubes.

4 Methodology

The two steps methodological approach proposed for this study can be summarized as follows: i) A predictive model of the different emission sources' contributions to the pollutant concentrations, captured at each monitoring station, divides the time series into two classes: pairs of highly correlated points and pairs of points with poor correlation; ii) Validation of a time model for the entire area.

4.1 CLASSIFICATION OF TIME PERIODS WITH HIGH CORRELATION BETWEEN EMISSION AND MONITORING STATION RECORDS

After the first attempts of including meteorological variables into the prediction models, we concluded that the available data of wind speed and direction wasn't responsible for the observed dynamics of the different pollutant plumes; the main reason being that the meteorological data was often collected at an altitude and locations inadequate to capture emissions from the industries' chimneys.

The data sets were grouped into data pairs in order to analyse possible relations between each emission and each reception. Each data pair is composed by one industrial emission and one monitoring station record. In a first step, the time series of each data pair was screened with the purpose of obtaining contiguous time periods with high correlation between a specific industrial emission and the equivalent monitoring station measurements. The selection process consists in the implementation of a simple iterative procedure. The variogram of each pair of emissions-monitoring station AQ measurements during a period T (365 days – N error values) is:

$$\gamma(z_1,\psi_1)=\frac{1}{T}\sum_{i=1}^{T}\left[z_1(i)-\psi_1(i)\right]^2 , \tag{1}$$

where $z_1(i)$ and $\Psi_1(i)$ are the measurements of the emission source z_1 and of the monitoring station Ψ_1 for the instant *i* after standardization. This variogram was assumed as an objective function that tends to decrease (increasing the correlation between z_1 and Ψ_1) as pairs of points with less contribution are iteratively removed.

The selection process separates the data points into 2 classes: Class 1 – pairs of points with high correlation between a specific emission and one reception; Class 2 – pairs of points with low correlation between a specific emission and one reception.

Afterwards, a probabilistic neural network (PNN) was used to automatically classify data into the two classes described above. PNNs can be useful for classification problems and have a straightforward design. A PNN is guaranteed to converge to a Bayesian classifier, providing it is given enough training data, and generalizes well (Haykin 1994, Beale and Demuth 1998).

4.2 VALIDATION OF TIME MODEL FOR THE ENTIRE AREA

The obtained PNN is a predictive (classification) model valid for a period with statistical characteristics identical to the past and for the emissions-AQ monitoring station records pairs. The objective of the proposed methodology's second step is to validate and generalize this classification model for the entire area. In other words, to analyse the spatial extension of the classification model, calculated and tested for the AQ monitoring stations.

Hence the following geostatistical methodology is applied:

i) First, diffusive tubes measurements are used to determine a local trend of the SO_2 concentration corresponding to the 11 days period, through ordinary kriging;

ii) Based on the diffusive tubes variograms (spatial pattern) and the monitoring stations AQ values, a set of simulated images of SO_2 is obtained for the 11 days period, using direct sequential simulation (Soares, 2000) with local means, i.e., the local trend previously calculated;

iii) To validate the classification model for the entire area, the individual contributions of different emissions are mapped as follows: After averaging the simulated images for each day, the resulting most probable image was classified with a PNN (*cf.* Section 4.1), resulting in areas with high and low correlation with the different emissions.

5 Results and Discussion

5.1 CLASSIFICATION OF TIME PERIODS WITH HIGH CORRELATION BETWEEN EMISSION AND MONITORING STATION RECORDS

With the purpose of obtaining contiguous time periods with high correlation between each pair of industrial emission and monitoring station measurements (class 1 data points), the time series of each data pair was previously partitioned using the methodology described in Section 4.1. An example scatter plot of the standardized values of the original data series, for the Petrogal (emission) and Sonega (monitoring station) pair, is shown in figure 6 (a). A scatter plot of Class 1 data points (high correlation between Petrogal and Sonega's records) is shown in figure 6 (b). Figure 7 represents the time series of these Class 1 values, showing a contiguous time period, *i.e*, time period where, in principle, the meteorological conditions are in accordance with the direction emission/AQ monitoring station.

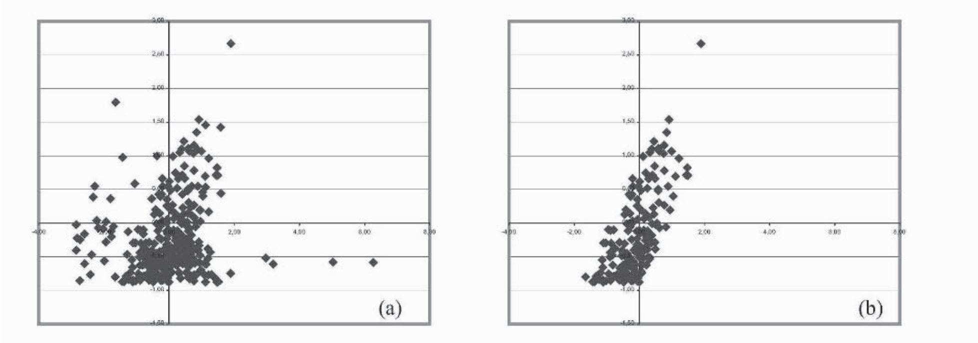


Figure 6. (a) Petrogal (x-axis) and Sonega's (y-axis) SO_2 concentrations before being selected; (b) Petrogal and Sonega's Class 1 data points.

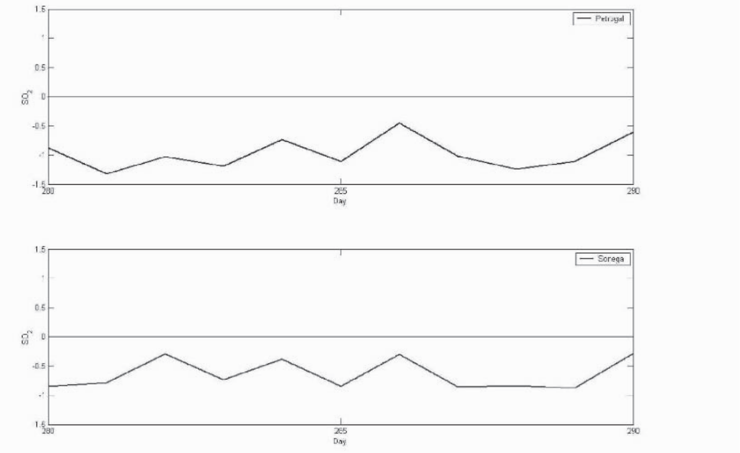


Figure 7. Example of contiguous Petrogal and Sonega's Class 1 data points.

The ability of the PNN to correctly classify time series into Class 1 and Class 2 exceeded 90%, for the three monitoring stations.

5.2 VALIDATION OF TIME MODEL FOR THE ENTIRE AREA

Validation of the PNN predictive model is necessary to evaluate the probability of other areas around the monitoring stations to belong to either of the classes of correlation with the emissions or, in other words, to evaluate the probability that the pollutant concentration in non-sampled locations is caused by the industrial emissions.
First, the diffusive tubes measurements (Fig. 5) were used to calculate (trough ordinary kriging) a local trend for the pollutant concentration for the 11 days period (Fig. 8).

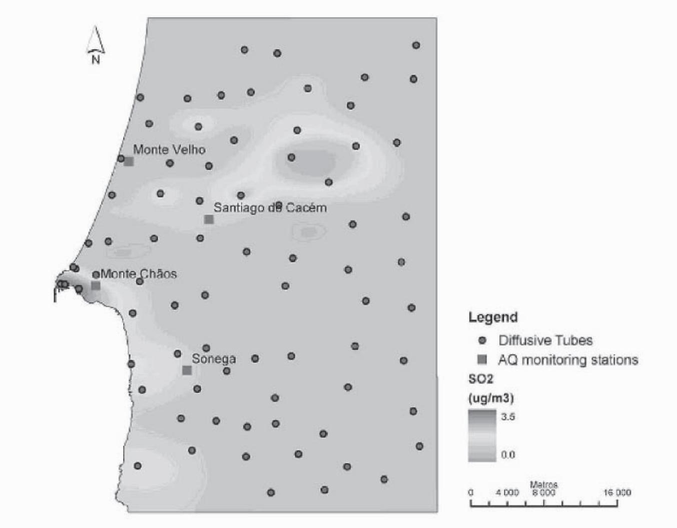


***Figure 8*.** Spatial trend of the SO_2 dispersion measured in the diffusive tubes.

As the diffusive tubes are the only available spatial data, it is assumed that the variogram calculated with this data reflects the spatial pattern of the average behaviour for the 11 days period. Hence, the variogram model for the diffusive tubes measurements – following an isotropic spherical model with one structure of range a=20 000 m – was considered for the subsequent steps.
Direct sequential simulation was applied to generate a set of 30 images. The local trend of figure 8 was assumed as local mean. AQ monitoring stations values of those 11 days were taken as conditioning data.
In figure 9 examples of SO_2 maps simulated for three consecutive days are shown. Average and variance maps for the first and last days of the 11 days period are shown in figures 10 a) and b), respectively.

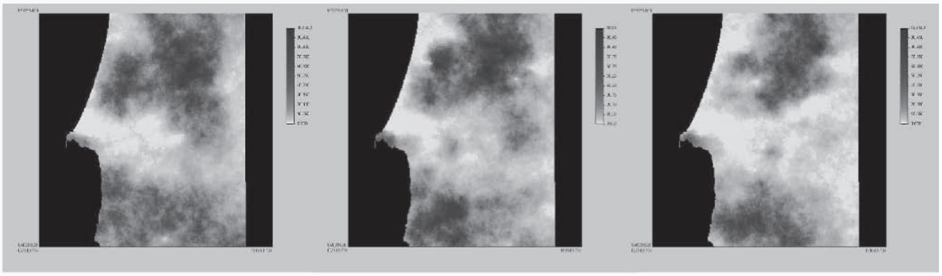


Figure 9. Examples of SO_2 maps simulated for three consecutive days.

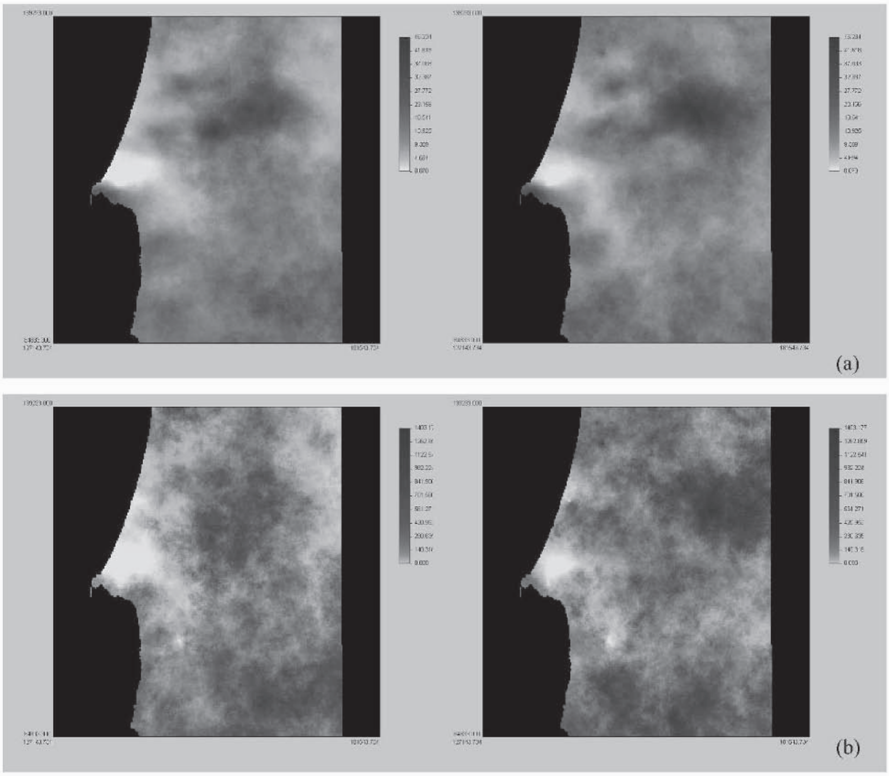


Figure 10. Examples of SO_2 average (a) and variance maps (b) for the first and last days of the 11 days period.

We attempted to calculate the correlation coefficient between each set of simulated images and each emission for the 11 days using the simulated spatial images. But, given the very homogeneous time period in terms of emissions, the resulting correlation coefficients constituted, most of the times, rather spurious statistics.

Hence, after averaging the simulated images for each day, the resulting most probable image was classified with a PNN (Fig. 11). Figure 11 shows the areas with highly correlated points and areas without correlation, with the different emissions.

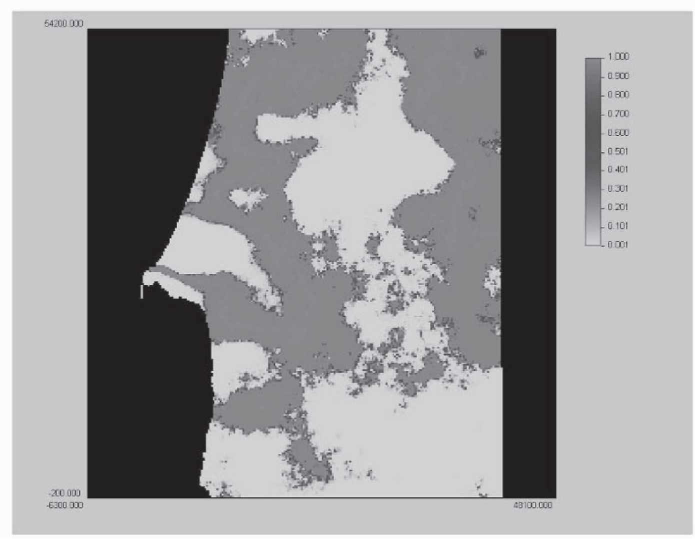


Figure 11. Areas with high and low correlation between emissions and the different receptions.

The PNN determines the probability of a given pair of points displaying a linear relationship between emissions and monitoring stations. All PNNs for the three industrial emissions and AQ monitoring stations are very similar, producing similar final maps for the entire region.

As all of the emissions, coincidently, show intermediate values for the 11 days period, one can see that:

i) The areas, which have a high probability of being related with the emissions, are the ones with intermediate values of pollutant concentration.

ii) The PNN determines the probability of a given pair of points belonging to the group of data displaying a linear relationship between emissions and monitoring station records. All PNN for the three industrial emissions and AQ monitoring stations are very similar. As the emissions are similar for the 11 days period of time, the final maps of each emission contribution to the pollution of entire region are also similar.

iii) The areas affected by high pollutant concentrations do not show any correlation with any of the industrial emissions. In fact, both hot-spots (high-value plumes) are located in the two main villages of the region, suggesting other pollutant source than the industrial emissions.

5.3 DISCUSSION

Combining two AQ sampling systems – classical monitoring stations and diffusive tubes – we succeeded in showing an approach that allows an impact evaluation of different emissions for the entire Sines area.

The predictive time model is strictly valid for the spatial location of the emissions/monitoring stations pairs. With a more spatially representative monitoring –

This page intentionally blank

using diffusive tubes – and with a spatial geostatistical model – through stochastic simulation – the model is successfully generalized for the entire area.

Inferences for the entire area are obviously just valid for the period of the diffusive tubes exposure. The more yearly campaigns of diffusive tubes become available, the more representative (in terms of space and time) the conclusions become.

In this case study, the conclusions drawn from the eleven days of the first campaign are just illustrative of the potential of the two steps approach. Although the results are coherent, the model is not validated for the entire space-time domain of the study.

6 Conclusions

This study deals with a well-known characteristic common to most AQ monitoring networks: high density of sample values in time, collected at just few spatial locations. This can be a serious limitation if one wishes to evaluate impact costs or carry out an environmental risk analysis of the emissions for the different land uses, eco-systems and natural resources of a region.

The presented approach, based on the use of two different monitoring systems – AQ monitoring stations, with an high density sampling rate in time, and diffusive tubes, that cover the entire space for a limited period of time – shows to be a valid alternative for an air-quality impact study covering the entire region.

In spite of the illustrative purpose of this paper, it is worth mentioning that the model should be validated for the entire area with more diffusive tubes campaigns. It is important to acknowledge that the model's performance could also be improved using longer AQ data series and another kind of meteorological data.

References

Beale, M.H. and Demuth, H.B., *Neural Network Toolbox for Use with MATLAB, User's Guide, version 3*. The MathWorks, Inc, 1998.

Cobourn, W.G., Dolcine, L., French, M. and Hubbard, M.C., *Comparison of Nonlinear Regression and Neural Network Models for Ground-Level Ozone Forecasting*. J. Air & Waste Manage. Assoc., 50, 1999-2009, 2000.

Cruz, L. and Campos, V., *Amostragem passiva de poluentes atmosféricos. Aplicação ao SO2*. Quim. Nova, Vol. 25, No. 3, 406-411, 2002.

Haykin, S., *Neural Networks: A Comprehensive Foundation*. Macmillan Pub, New York, 1994.

Kolehmainen, M., Martikainen, H. and Ruuskanen, J., *Neural networks and periodic components used in air quality forecasting*. Atmospheric Environment, 35, 815-825, 2000.

Mukerjee, S., Smith, L.A., Norris, G.A., Morandi, M.T., Gonzales, M., Noble, C.A., Neas, L. and Ozkaynak, A.H., *Field Method Comparison between Passive Air Samplers and Continuous Monitors for VOCs and NO2 in El Paso, Texas*. J. Air & Waste Manage. Assoc., 54, 307–319, 2004.

Russo, A., Nunes, C. and Bio, A., *Air quality models resulting from multi-source emissions*, geoENV 2004 - Fifth European Conference on Geostatistics for Environmental Applications. Neuchâtel, Switzerland, 2004.

Soares, A., *Geoestatística Aplicada às Ciências da Terra e do Ambiente*, IST PRESS, 206p, 2000.

WHO, *Ambient Air Quality Monitoring and Assessment (Guidelines for Air Quality)*, World Health Organization, Geneva, 1999.

MAPPING LAND COVER CHANGES WITH LANDSAT IMAGERY AND SPATIO-TEMPORAL GEOSTATISTICS

ALEXANDRE BOUCHER, KAREN SETO and ANDRÉ JOURNEL
Department of Geological & Environmental Sciences, Stanford University, Stanford, CA, 94305-2115

Abstract.

Satellite images are the principal medium to detect and map changes in the landscape, both in space and time. Current image processing techniques do not fully exploit the data in that they do not take simultaneously into account the spatial and the temporal relations between the various land cover types. The method proposed here aims to accomplish that.

At each pixel of the landscape, the time series of land cover type is modeled as a Markov Chain. That time series at any specific location is estimated jointly from the local satellite information, the neighboring ground truth land cover data, and any neighboring previously estimated time series deemed well-informed by the satellite measurements.

The method is applied to detect anthropogenic changes in the Pearl River Delta, China. The prediction accuracy of the time series improves significantly, the accuracy almost double, when both spatial and temporal information are considered in the estimation process. The introduction of spatial continuity through indicator kriging also reduced unwanted speckles in the classified images, removing the need for post-processing.

1 Introduction

Remote sensing data and applications are fundamentally spatial in nature, yet the current image processing methods too often do not consider the spatial context when estimating the label of any given pixel. A common method to incorporate spatial information is to model the spatial distribution of labels as a Markov random field (Tso and Mather, 2001). However, the estimation call for an iterative algorithm which can be computationally demanding for large domains. Geostatistics, especially indicator kriging, has also been used to incorporate spatial autocorrelation in estimating or simulating labels (Atkinson and Lewis, 2000; Brown et al., 2002; Wang et al, 2004).

When pursuing change detection, that is the mapping in both space and time of land covers changes, a good cross-sectional accuracy is not enough, one must also

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 809–818.

ensure accuracy through time. The temporal component becomes as important as the spatial component when one desires to know both when and where changes have occurred.

The proposed methodology presents a way to integrate the spatial correlation of the land covers labels with temporal information thus improving the mapping of land cover changes.

The framework is applied to mapping anthropogenic changes in the Pearl River Delta, China. This region is going through tremendous growth in population with important environmental repercussion on the landscape, such as deforestation and urban sprawl.

2 Notations

Consider a domain $D \subset \mathcal{R}^2$ measured at different times $t_i, i = 1, ..., N_t \subset T$. The N_t measurements of D constitute a set of images denoted $\mathcal{I} = \{\mathcal{I}^{(t_1)}, ..., \mathcal{I}^{(t_{N_t})}\}$. Let $(\mathbf{u}, t)$ be a point in $D \times T$ informed by a vector of length n_B of continuous attributes, $\mathbf{Z}(\mathbf{u}, t) = \{Z_1(\mathbf{u}, t), ..., Z_{n_B}(\mathbf{u}, t)\}$. These attributes are the satellite measurements known as digital numbers (DN).

Each pixel $(\mathbf{u}, t)$ must be classified into one of K labels $\mathcal{L}_1, ..., \mathcal{L}_K$, for example K land cover types. Define $I_k(\mathbf{u}, t)$ an indicator variable indicating whether or not the pixel at location $(\mathbf{u}, t)$ has label $\mathcal{L}_k$

$$I_k(\mathbf{u}, t) = \begin{cases} 1 & \text{if } \ (\mathbf{u}, t) \in \mathcal{L}_k \\ 0 & \text{otherwise} \end{cases}$$

And let

$$\mathcal{L}(\mathbf{u}, t) = k \quad \text{if } I_k(\mathbf{u}, t) = 1$$

Furthermore, let Ω be the set of location $\mathbf{u}_\alpha, \alpha = 1, ..., n$ whose labels are known at all times (ground truth). $V(\mathbf{u}, t)$ is the set of known labeled pixel data in an isochronous neighborhood of $\mathbf{u}$ at time t.

3 Coding and combining information

The available information at each uninformed location $\mathbf{u}$ is first separated between isochronous (cross-sectional) and time series information. The isochronous or cross-sectional information includes the satellite response and the neighboring land cover indicators at a specific time, the time series information consist of transition probabilities linking the land cover indicators through time. The classification at location $\mathbf{u}$ is then done by combining these two types of information in such a way to minimize misclassification over a given training set.

3.1 TIME SERIES TRANSITION PROBABILITIES

Denote by $p_k^{\mathrm{T}}(\mathbf{u},t)$ the probability of having label $\mathcal{L}_k$ at location $(\mathbf{u},t)$ given the collocated land cover indicators in the immediate past $(\mathcal{L}(\mathbf{u},t-\Delta_1 t))$ or future $(\mathcal{L}(\mathbf{u},t+\Delta_2 t))$, or in both past and future.

$$p_k^{\mathrm{T}}(\mathbf{u},t) = \mathrm{Prob}\{I_k(\mathbf{u},t)=1 \mid \mathcal{L}(\mathbf{u},t-\Delta_1 t), \mathcal{L}(\mathbf{u},t+\Delta_2 t)\ \} \tag{1}$$

The probability $p_k^{\mathrm{T}}(\mathbf{u},t)$ is calibrated directly from ground truth data or determined as function of the transition probabilities $p_{kk'}(t_i,t_j)$ relating the probability of having class $\mathcal{L}_{k'}$ at time t_j given that $\mathcal{L}_k$ is observed at time t_i.

$$p_{kk'}(t_i,t_j) = \mathrm{Prob}\{I_{k'}(\mathbf{u},t_j)=1 \mid I_k(\mathbf{u},t_i)=1\}, \forall\ \mathbf{u},k,k' \tag{2}$$

The transition probabilities $p_{kk'}(t_i,t_j)$ are calibrated from ground truth data.

3.2 ISOCHRONOUS PROBABILITIES

The isochronous information at any specific time is obtained by combining the satellite response and the spatial information available at that time. All information is expressed in terms of probabilities. Denote by $p_k^{\mathrm{iso}}(\mathbf{u},t)$ the isochronous probability obtained by combining the probabilities $p^{\mathrm{DN}}(\mathbf{u},t)$ and $p^{\mathrm{S}}(\mathbf{u},t)$ obtained from the satellite and spatial information respectively.

$$\begin{aligned} p_k^{\mathrm{iso}}(\mathbf{u},t) =&\ \mathrm{Prob}\{I_k(\mathbf{u},t)=1 \mid \mathbf{Z}(\mathbf{u},t), \mathcal{L}(\mathbf{u}',t), \mathbf{u}' \in V(\mathbf{u},t)\} \\ =&\ \phi(p_k^{\mathrm{DN}}(\mathbf{u},t), p_k^{\mathrm{S}}(\mathbf{u},t)) \end{aligned} \tag{3}$$

The combination algorithm ϕ is presented later.

Satellite-derived probabilities

The conditional probability $p_k^{\mathrm{DN}}(\mathbf{u},t)$ for the pixel at location $(\mathbf{u},t)$ to be assigned to label $\mathcal{L}_k$ given the satellite response is computed with a classifier $F(\cdot)$ calibrated from the known data $\{\mathbf{Z}(\mathbf{u}_\alpha,t), \mathcal{L}(\mathbf{u}_\alpha,t)\}$ (Richards and Jia , 1999). The function $F(\cdot)$ approximates the conditional expectation of $I_k(\mathbf{u},t)$ given the sole collocated satellite response.

$$p_k^{\mathrm{DN}}(\mathbf{u},t) =\ \mathrm{Prob}\{I_k(\mathbf{u},t)=1 \mid \mathbf{Z}(\mathbf{u},t)\},\ \forall\ k \tag{4}$$

In this study, the conversion of Landsat TM measurements into land cover types probabilities is done with the conventional maximum likelihood (ML) classifier (Richards and Jia , 1999). The principle is simple, the probabilities $\mathrm{Prob}\{I_k(\mathbf{u},t) = 1 \mid \mathbf{Z}(\mathbf{u},t)\}, k=1,..,K$ are calculated from the training set using a Bayes' inversion

$$p_k^{\mathrm{DN}}(\mathbf{u},t) = \mathrm{Prob}\{I_k(\mathbf{u},t)\ =1|\ \mathbf{Z}(\mathbf{u},t)=\mathbf{z}\} = \frac{\mathrm{Prob}\{\mathbf{Z}(\mathbf{u},t)=\mathbf{z}|I_k(\mathbf{u},t)=1\}\mathrm{Prob}\{I_k(\mathbf{u},t)=1\}}{\sum_{k'=1}^{K}\mathrm{Prob}\{\mathbf{Z}(\mathbf{u},t)=\mathbf{z}|I_{k'}(\mathbf{u},t)=1\}\cdot\mathrm{Prob}\{I_{k'}(\mathbf{u},t)=1\}}$$

Assuming the random vector $\mathbf{Z}(\mathbf{u},t)$ to be multiGaussian, its conditional probability is written as

$$\text{Prob}\{\mathbf{Z}(\mathbf{u}) = \mathbf{z}|I_k(\mathbf{u},t) = 1\} = \frac{1}{(2\pi)^{N/2}|\Sigma_k|^{1/2}} e^{-\frac{1}{2}(\mathbf{z}-\mathbf{m}_k)^T \Sigma_k^{-1}(\mathbf{z}-\mathbf{m}_k)} \quad (5)$$

where $\mathbf{m}_k$ and Σ_k are the mean vector and covariance matrix of the DN values belonging to the training data with label $\mathcal{L}_k$.

Spatially-derived probabilities
Denote by $p_k^{\mathrm{S}}(\mathbf{u},t)$ the conditional probability of observing $\mathcal{L}_k$ at location $(\mathbf{u},t)$ given the isochronous label data found in the neighborhood $V(\mathbf{u},t)$.

$$p_k^{\mathrm{S}}(\mathbf{u},t) = \text{Prob}\{I_k(\mathbf{u},t) = 1 \mid \mathcal{L}(\mathbf{u}',t), \mathbf{u}' \in V(\mathbf{u},t)\}, \ \forall\ k \quad (6)$$

This spatial probability $p_k^{\mathrm{S}}(\mathbf{u},t)$ may be estimated from simple indicator kriging (Goovaerts, 1997). Simple indicator kriging is a linear interpolator that applied kriging weights to indicator data yielding the probability of belonging to a class given the neighborhood data, the marginal and the covariance model of that class.

In addition to the ground truth data, the neighboring data in $V(\mathbf{u},t)$ also include locations that are considered well informed by the sole satellite measurements. Any measure of information content could be used to determine which locations are well informed and which ones are not. Those well-informed nodes are locations where the DN measurements $\mathbf{Z}$ alone are deemed sufficient to label them. For example, a pixel where the classifier $F(\cdot)$, see expression (5), would indicate a probability of 0.98 or more to belong to a certain label would qualified as a well-informed node.

Those nodes, assumed to be fully informed by the sole satellite information, are used as anchor for the less informed ones. This spreads information from high-confidence pixels to their surrounding. For example, if in an area all the well-informed locations are urban, the neighboring pixel are more likely to belong to the urban label. The indicator kriging from the well-informed labels performed just that.

There is however a risk to overextend the spatial relevance of the well-informed locations. The problem lies in the discontinuity of the landscape. For example a certain region may be predominantly urban, without forest or agriculture, but the vegetated area could start abruptly a few pixels away. A well-informed water label located in a lake close to the shore does not say whether that shore is urbanized or vegetated, instead it tends to artificially increase the probability that the shore would belong to a water label.

To offset this problem of borders and discontinuities, the images are first segmented to find edges delineating those discontinuities. Then a data neighborhood that does not cross the edges is retained for the indicator kriging process. Interpolation (kriging) is thus limited to homogeneous neighborhoods.

3.3 POSTERIOR PROBABILITY

The posterior probability $p_k(\mathbf{u},t)$ for class $\mathcal{L}_k$ to occur at location $(\mathbf{u},t)$ is computed by combining the isochronous probability $p_k^{\text{iso}}(\mathbf{u},t)$ and the time series probability $p_k^{\text{T}}(\mathbf{u},t)$

$$p_k(\mathbf{u},t) \;=\; \text{Prob}\{I_k(\mathbf{u},t)=1 \mid \text{all data}\} \;=\; \psi(\; p_k^{\text{iso}}(\mathbf{u},t), p_k^{\text{T}}(loc,t)\;),\; \forall\, k$$

The proposed combination algorithm ψ is developed in the next section.

Finally, the label $\mathcal{L}(\mathbf{u},t)$ is estimated by taking the most probable class from the posterior distribution such that

$$\mathcal{L}^*(\mathbf{u},t) = \arg\max_k \;\{p_k(\mathbf{u},t),\; k=1,..,K\} \tag{7}$$

The time series, $\{\mathcal{L}(\mathbf{u},t_1),...,\mathcal{L}(\mathbf{u},t_{N_t})\}$ at location $\mathbf{u}$ is generated by estimating the labels starting from the most informed and then sequentially estimating the time before and after that starting time. The idea is that the starting point is very consequential for the estimation of the whole time series, that starting time is thus chosen to reduce the prediction error. The less informed times at any given location would benefit from being conditioned on the better informed collocated times.

3.4 COMBINING PROBABILITIES

Consider the isochronous probability vector $p^{\text{iso}}(\mathbf{u},t)$ defined in expression (3) and the time series conditional probability $p_k^{\text{T}}(\mathbf{u},t)$ defined in expression (1) as two sources of information. Each of the those two probabilities can be transformed into a distance related to the likelihood of event $\mathcal{L}(\mathbf{u},t)=k$ occurring (Journel, 2002). Let that distance[1] be

$$x_{\mathcal{L}_k}^{\text{iso}}(\mathbf{u},t) = \frac{1-p_k^{\text{iso}}(\mathbf{u},t)}{p_k^{\text{iso}}(\mathbf{u},t)} \in [0,\infty]$$

$$x_{\mathcal{L}_k}^{\text{T}}(\mathbf{u},t) = \frac{1-p_{\mathcal{L}_k}^{\text{T}}(\mathbf{u},t)}{p_{\mathcal{L}_k}^{\text{T}}(\mathbf{u},t)\}}$$

Consider also the distance related to the marginal probabilities

$$x_{\mathcal{L}_k}^{(0)} = \frac{1-\text{Prob}\{(\mathbf{u},t)\in\mathcal{L}_k\}}{\text{Prob}\{(\mathbf{u},t)\in\mathcal{L}_k\}}\;\forall\mathbf{u}$$

The updated distance to the event $\mathcal{L}(\mathbf{u},t)=k$ occurring accounting for both information (1) and (3) is given by the "tau model":

$$x_{\mathcal{L}_k}(\mathbf{u},t) = x_{\mathcal{L}_k}^{(0)}\cdot\left(\frac{x_{\mathcal{L}_k}^{\text{iso}}(\mathbf{u},t)}{x_{\mathcal{L}_k}^{(0)}}\right)^{\tau_{\text{iso}}}\cdot\left(\frac{x_{\mathcal{L}_k}^{\text{T}}(\mathbf{u},t)}{x_{\mathcal{L}_k}^{(0)}}\right)^{\tau_{\text{T}}} \tag{8}$$

[1] Note that the distances are the inverse of odd-ratio used in logistic regression.

where τ_{iso} and τ_{T} are parameters introducing redundancy between the two information sources (Journel, 2002; Krishnan et al, 2004). This study assume all τs equal to 1, corresponding to conditional independence between the two sources. The posterior probability is then retrieved by

$$\mathrm{Prob}\{(\mathbf{u},t) \in \mathcal{L}_k | \mathbf{Z}(\mathbf{u},t)\} = \frac{\frac{1}{1+x_{\mathcal{L}_k}(\mathbf{u},t)}}{\frac{1}{1+x_{\mathcal{L}_1}(\mathbf{u},t)} + \frac{1}{1+x_{\mathcal{L}_2}(\mathbf{u},t)} + \ldots + \frac{1}{1+x_{\mathcal{L}_K}(\mathbf{u},t)}} \tag{9}$$

The integration of $p^{\mathrm{DN}}(\mathbf{u},t)$ and $p^{\mathrm{S}}(\mathbf{u},t)$ into $p^{\mathrm{iso}}(\mathbf{u},t)$ is also done with expression (9) but using different tau parameters τ_{S} and τ_{DN}.

4 A case study, urbanization in the Pearl River Delta, China

The Pearl River Delta in China has seen its population soar with economic development in the last three decades. The region under study is centered on the city of Shenzhen, in the Guangdong province. Anthropogenic changes are important and Landsat imagery has already been used to map the changes in the landscape (Kaufmann and Seto, 1989; Seto et al., 2002).

This study focused on detecting and mapping changes that has happened between between 1988 and 1996 using a time series of Landsat 7TM images. We acquired 6 images dating from 1988,1989, 1992, 1994, 1995 and 1996 all taken around December. The year 1990 and 1991 are not used because of significant cloud cover, while 1993 was let aside because of poor georeferencing. The Band 7 for 1988 and 1996 is shown in Figure 1. There are 1917870 times series informed by the satellites approximatively covering an area of size 45km by 45 km, with each pixel of dimension 30x30 meters.

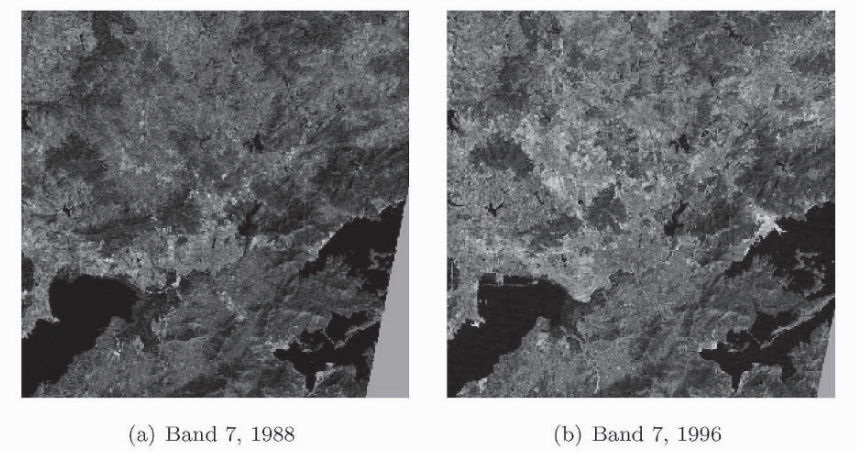
The landscape is divided into K=7 classes: water, forests, agriculture, urban, fish pond, transition (land getting cleared for urban settlement) and shrub. The ground truth measurements consists of 1917 locations identified by expert interpretation or by field reconnaissance. At ground truth locations the labels are deemed known at all times. The prediction errors are estimated by a 5-fold cross-validation procedure (Hastie et al., 2001). The known labels are divided five times, each time into a training set and a testing set such that all samples are used once both for testing purposes. Each split is done such that 80% of the ground truth data belong to the training set and the remainder 20% to the test set.

4.1 COMPUTING THE TRANSITION PROBABILITIES

The time series transition probabilities $p_{kk'}(t_i,t_j)$ defined in expression (2) are assumed stationary in time, such that

$$p_{kk'}(t_i,t_j) = p_{kk'}(\Delta t)$$

where $\Delta t = t_j - t_i$. The $p_{kk'}(\Delta t)$ are computed from the training set by evaluating the proportions of transitions from class k to class k'. Notable characteristics of this transition probability matrix is that the urban land cover type is an absorbing


(a) Band 7, 1988

(b) Band 7, 1996

Figure 1. Band 7 over the study area for 1988 and 1996. The black pixels are water, the gray areas are agriculture and forest land covers while the bright spots are mostly urban and transition land covers. Note the larger urban area in 1996 than in 1988.

state while the transition land cover type only communicates with itself and with the urban land cover state. This means that once a pixel is urban, it will remain urban; and if a pixel has a transition label, it can remain in transition or become urban.

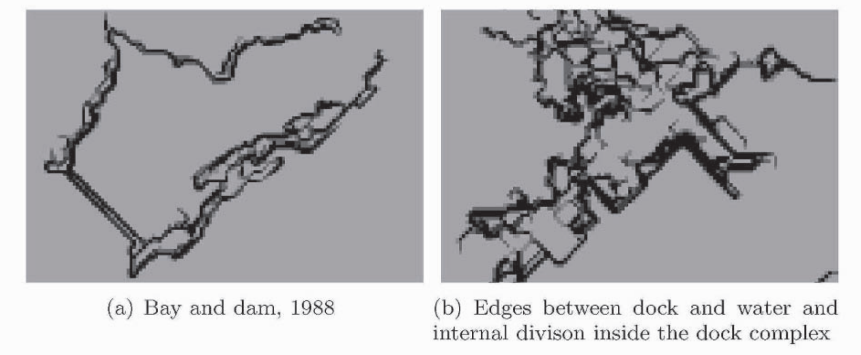
4.2 COMPUTING ISOCHRONOUS PROBABILITIES

Satellite-derived probabilities
The probabilities $p_k^{\mathrm{S}}(\mathbf{u}, t), k = 1, ..., K$ is computed with a maximum likelihood estimator, see expression (5).

Spatially-derived probabilities
The spatial context is accounted for through the probabilities $p^{\mathrm{S}}(\mathbf{u}, t)$ estimated with simple indicator kriging using for conditioning data the time series at locations deemed well informed by the satellite measurements. The information content of a time series at location $(\mathbf{u})$ is measured as the sum of the maximum satellite-derived probability at each times.

$$\mathrm{Inf}(\mathbf{u}) = \frac{1}{N_t} \sum_{i=1}^{N_T} \max(p_k^{\mathrm{DN}}(\mathbf{u}, t_i), k = 1, ..., K) \tag{10}$$

After some trials and errors, a time series is deemed well informed if $\mathrm{Inf}(\mathbf{u})$ is greater than 0.87. Furthermore, those well informed time series will only be included in the neighborhood if a straight line going from the center of the neighborhood to any well informed datum does not cross an edge. The edges are found by performing a Canny segmentation method (Canny, 1986). Figure 2 shows two examples of edge detection. The edges in Figure 2(a) represent the shore of a bay with a dam at the western extremity. In Figure 2(b), the edges delineate a port from the ocean and also segment homogeneous region inside the port complex.



(a) Bay and dam, 1988

(b) Edges between dock and water and internal divison inside the dock complex

Figure 2. Examples of edge detection. In Figure (a), the edges capture the border of the bay and the dam at its extremity. In Figure (b), the edge define the contact between a port and the bay plus some internal divisions inside the port complex.

4.3 RESULTS

The results of the proposed method are compared to the accuracy resulting from the maximum likelihood (ML) classifier, see expression (5). The ML classification is done by assigning to a time-space location $(\mathbf{u}, t)$, the class that has the maximum probability $p_k^{\mathrm{DN}}(\mathbf{u}, t)$. This classification solely considers the satellite responses thus ignoring the temporal and spatial correlation between labels.

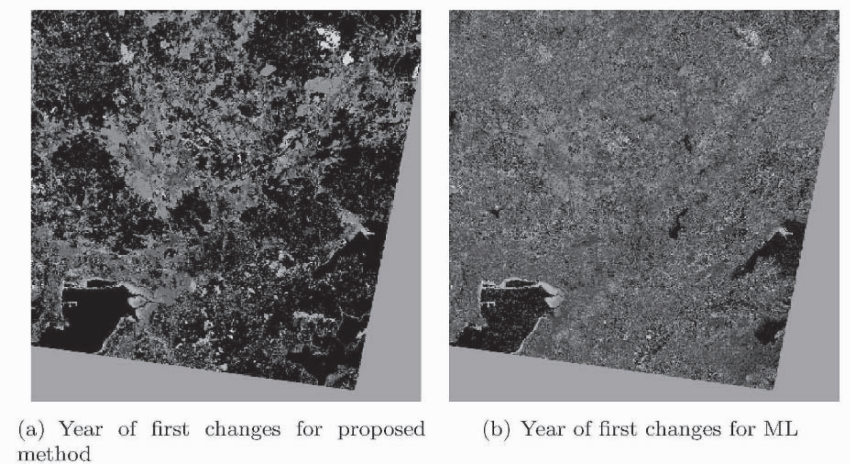
The results are validated using (1) the overall accuracy ,the percent of correctly classified pixels, and (2) the time series accuracy, the percent of locations which have their vector of labels **all** correct. A time series at location $\mathbf{u}$ is well classified only if its six labels have been correctly predicted. For change detection purpose, the time series accuracy is important as it shows how well the changes are mapped in time and space.

With the ML classifier, the accuracy from the five-fold cross validations yields an overall accuracy of 78%, but the time series accuracy drops to 33%. The proposed method only marginally improves the overall accuracy from 78% to 82%. However, the accuracy of the time series goes up 61%, a considerable improvement.

The indicator kriging also decreases the level of speckling in the images, producing smoother maps. For example, the ML tend to classify many shadow zones in mountainous areas as water, the integration of spatial information corrects many of those misclassified pixels. This no need to post-processed the classified images to remove the speckles.

The maps in Figure 3 show, for each location, the year at which change first occurred. A comparison between Figure 3(a) and (b) clearly shows that the proposed method preserves some spatial relationships for the land cover changes, exhibiting a structured evolution of the landscape. On the contrary, the ML method produces a salt and pepper texture where the physical evolution of the landscape is indiscernible.

The proposed method also provides a more stable and more realistic mapping of the changes. With the ML prediction, 35% of locations had changed more than once, a number that visual inspection of the images and knowledge about the area do not validate. Only 9% of location are predicted to change more than once with the proposed method. The ML also predicts that 22% of locations did not change while that percent goes up to 64% when the spatial and temporal information are combined in the prediction.


(a) Year of first changes for proposed method

(b) Year of first changes for ML

Figure 3. Map of predicted land cover changes representing the year at which the first change occurred. Figure (a) maps the year of change as predicted by the proposed method. Figure (b) does it for the ML method. Black indicates no changes, lighter tones indicates later times. Note the greater spatial resulution for the proposed method.

5 Conclusion

This paper proposed a framework that allows the integration of the spatial and temporal autocorrelation of labels in remote sensing applications. That integration produces a more accurate change detection map that better defines when and where the landscape had changed. This study uses 6 images, the extension to longer time series would be straightforward as the complexity of the algorithm only increases linearly with additional images.

Prior identification of well-informed locations from the sole satellites information is shown to work well. Enough conditioning data are made available so that geostatiscal methods can be used in an estimation mode to improve the land covers estimation.

Importantly, the study shows a considerable increase of time series accuracy with the proposed method. Furthermore, the evolution of the landscape display greater spatial continuity and seems more realistic.

Future works will focus on improving the method by using more complex and potentially better suited spatial model such as training images replacing variogram models; as well as better way to incorporate the time series information. Another avenue of research is the modeling of the tau parameters in the combination algorithm (8), to weight each individual source of information and further increase the performance of the process.

References

Atkinson, P. M., and Lewis, P., *Geostatistical Classification for Remote Sensing: an Introduction*, Computers & geosciences, vol. 26, no. 4, 2000, 361-371

Brown, D. G., Goovaerts, P., Burnicki, A., and Li, M. Y., *Stochastic Simulation of Land-Cover Change using Geostatistics and Generalized Additive Models*, Photogrammetric engineering and remote sensing, vol. 68, no. 10, 2002, p. 1051-1061

Canny, J., *A Computational Approach To Edge Detection*, A Computational Approach To Edge Detection, vol. 8, p. 679-714, 1986

Goovaerts, P., *Geostatistics for Natural Resources Evaluation*, Oxford University Press, 1997.

Hastie, T., Tibshirani, R. and Friedman, J., *The Element of Statistical Learning*, Springer, 2001.

Journel, A.G., *Combining Knowledge from Different Sources: an Alternative to Conditional Independence Hypothesis*, Math. Geol., vol. 34, no. 5, 2002, p. 573-596

Kaufmann,R.K. and Seto K.C., *Change Detection, Accuracy and Bias in a Sequential Analysis of Landsat Imagery in the Pearl River Delta, China: Econometric Techniques*, Agriculture, Ecosystems and Environment, vol. 85, 2001, p. 95-105.

Richards, J. A. and Jia X. *Remote Sensing Digital Image Analysis, 3rd Edition*, Springer-Verlag, Berlin, 1999.

Seto, K.C., Woodcock, C.E., Song, C., Huang, X., LU, J. and Kaufmann, R.K., *Monitoring Land-use Change in the Pearl River Delta using Landsat TM*, Int. J. Remote sensing, vol. 23, no. 10, 2002, p. 1985-2004.

Krishnan, S., Boucher, A., and Journel, A.G., *Evaluating Information Redundancy Through the Tau Model*, In Proceedings of Geostatistical Congress 2004, Banff, Canada

Tso, B. and Mather, P.M. *Classification Methods for Remotely Sensed Data*, Taylor and Francis, London, 2001

Wang, G., Gertner,G., Fang, S. and Anderson, A.B., *Mapping Vegetation Cover Change using Geostatistical Methods and Bitemporal Landsat TM Images*, IEEE Transactions on Geoscience and remote sensing, vol. 42, no. 3, 2004, p. 632-243.

SPHERICAL WAVELETS AND THEIR APPLICATION TO METEOROLOGICAL DATA

HEE-SEOK OH
Department of Statistics, Seoul National University, Seoul, Korea
and
Department of Mathematical & Statistical Sciences, University of Alberta, Edmonton, Canada

Abstract. Li (1999) proposed the multiscale spherical wavelet (SW) method. In this paper, we investigate the potential of the multiscale SW representation for a climate field through a real experimental study. To represent a field by the multiscale SW method effectively, the appropriate choices of networks $\mathcal{N}_l$ and spherical basis function (SBF) $G_l(x)$ are required. We discuss some practical schemes to choose $\mathcal{N}_l$ and $G_l(x)$.

1 Introduction

Spherical data can be found in many applications such as atmospheric sciences where observations can be regarded as taken at different locations on a unit sphere. The data set used in this paper is a Northern Hemisphere winter mean temperature that can be obtained from the average of December, January and February raw temperatures. They selected 1000 stations to cover the whole sphere for the period of 1961-1990. Figure 1 shows the distribution of the 1000 stations used in our analysis. This data set was taken from a large database prepared by Jones et al. (1991). As shown in Figure 1, the data are scattered: they are not observed on regular spatial grids, and they have nonhomogeneous spatial densities including data voids of various sizes.

Given the scattered observations, one important problem is to represent the underlying spherical temperature field $T(n)$ for every location n on the globe. To represent $T(n)$, one could use the classical methods of spherical harmonics and spherical smoothing splines. However, as expected from the drawbacks of Fourier series estimators on the real line, spherical harmonics are not efficient for representing nonhomogeneous fields. The smoothing spline method tends to produce uniformly smooth results, although the data have intrinsic multiscale structure. Notice that the estimate of smoothing splines can be considered as the kriging estimates of Gaussian random field in spatial statistics (Cressie, 1993). The multiscale SWs proposed by Li (1999), on the other hand, are endowed with

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 819–831.

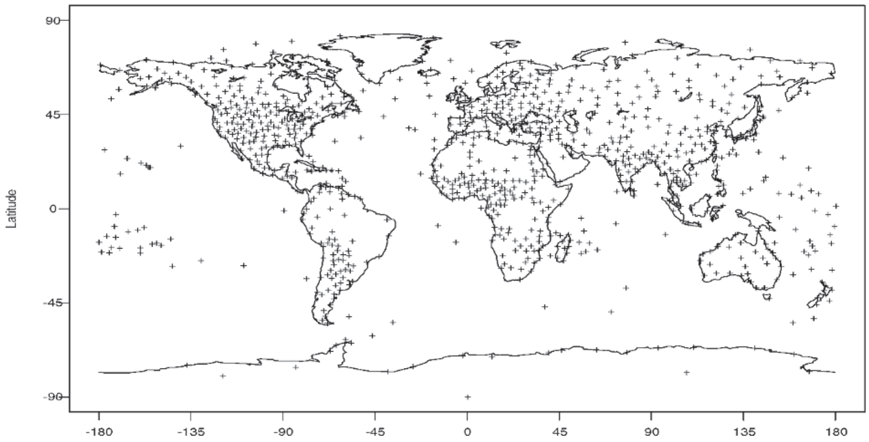


Figure 1. The distribution of the 1000 stations.

localization properties and therefore are particularly effective in representing multiscale phenomena that comprise activities of different scales at different locations. Moreover, the orthogonality of wavelets gives rise to multiscale decompositions that make the wavelet method a powerful tool for extracting the field's activities at different scales and detecting regional anomalies from global trends

However to representation a spherical field by using SWs effectively, it is required to decide some important issues in practice such as the appropriate choice of networks $\mathcal{N}_l$, SBF $G_l(x)$ and the coefficients $\boldsymbol{\beta}_1$. We propose methods to choose the practical issues, and apply multiscale SW method coupled with appropriate choices to data in Figure 1.

The rest of the paper is organized as follows. Section 2 gives a brief discussion the SW method of Li (1999). In section 3, we discuss some practical issues with respect to the implementation of the method. The results for the surface air temperature data are illustrated in Section 4.

2 Multiscale spherical wavelets

In this section, we briefly review the SW method proposed by Li (1999).

2.1 MULTISCALE SBF REPRESENTATION

The general idea of spherical basis function (SBF) representation is to use localized function call SBF to approximate an integrated squared underlying function on sphere. A detailed information of SBF can be found in Narcowich and Ward (1996). Under the spherical coordinate system, $n := [\cos\phi\cos\theta, \cos\phi\sin\theta, \sin\phi]^T$ denotes the unit vector that points to a location from the center of the unit sphere, where ϕ and θ denote the latitude and longitude of the location respectively. For a given network of J observation stations $\mathcal{N}_1 := \{n_j\}_{j=1}^{J}$, let us assume that we have created a nested sequence of networks $\mathcal{N}_1 \supset \mathcal{N}_2 \supset \cdots \supset \mathcal{N}_L$ and have associated the subnetwork $\mathcal{M}_l := \mathcal{N}_l \backslash \mathcal{N}_{l+1}$ with a spherical basis function (SBF) $G_l(x)$ which may have a different bandwidth for different values of l, where x is the cosine of the angle between two locations on the sphere. A specific example of constructing the networks and selecting the bandwidth will be given in Section 4. In this way, the original network $\mathcal{N}_1$ is partitioned into L subnetworks of different scales: $\mathcal{N}_1 = \bigcup_{l=1}^{L} \mathcal{M}_l$, and each subnetwork is given an SBF with a different bandwidth. For convenience, let us relabel the stations n_j using a double subscript notation such that $\mathcal{M}_l = \{n_{lj}, j = 1, \ldots, M_l\}$, where $\sum_{l=1}^{L} M_l = J$.

An SBF, $G(x)$, is a square-integrable and rapidly decaying function of $x \in [-1, 1]$ whose coefficients in the Legendre series expansion are all positive except a finite many that can be zero (Freeden et al., 1998). A simple and useful example is the Poisson kernel,

$$G(x; \eta) := \frac{1 - \eta^2}{(1 - 2\eta x + \eta^2)^{3/2}} = \sum_{m=0}^{\infty} (2m + 1)\eta^m P_m(x), \tag{1}$$

where $\eta \in (0, 1)$ is a bandwidth parameter and $P_m(x)$ is the Legengre polynomial of degree m. Note that the normalized Poisson kernel $\widetilde{G}(x; \eta) := \frac{(1-\eta)^2}{(1+\eta)} G(x; \eta)$ satisfies $\widetilde{G}(0; \eta) = 1$. As can be seen, the Poisson kernel has a peak at $x = 0$ and decreases monotonically as x deviates from 0 to ± 1. The bandwidth of the Poisson kernel is small when η is large, and the bandwidth is large when η is small.

A multiscale SBF representation of $T(n)$ takes the form of

$$T_1(n) = \sum_{l=1}^{L} \sum_{j=1}^{M_l} \beta_{1j} G_l(n \cdot n_{lj}), \tag{2}$$

where the dot product $n \cdot n_{lj}$ is equal to the cosine of the angle between n and n_j at subnetwork l. In this expression, the SBFs $G_l(x)$ have different bandwidths, η_l according to scale index l. Note that the SBFs at a fixed subnetwork l in (2) depend only on the angles between the location n and the observation sites n_j, so that the SBF representation is invariant to any rotations of the spherical coordinate system. In this representation, a large bandwidth is allowed for sparsely located stations and a small bandwidth for densely located stations. Furthermore, the nested networks $\mathcal{N}_l$ can be arranged so that the sparseness of stations in $\mathcal{N}_l$ increases with the increase of l. One can also choose the SBFs so that the bandwidth of $G_l(x)$ increases with the sparseness of $\mathcal{N}_l$, and thus the variable l can be truly regarded as a scale parameter. This can be accomplished, for example, by using the bottom-up design (BUD) procedure discussed in Li (2001).

Now let us describe a decomposition procedure that decomposes the SBF representation (2) into global and local components. For any given $l = 1, \ldots, L$, let

$$\mathcal{V}_l := \text{span}\{G_k(n \cdot n_{kj}) : \; j = 1, \ldots, M_k; \; l \leq k \leq L\}$$

be the linear subspace of all SBFs in (2) whose scales are greater than or equal to l. Define the inner product of two spherical fields by $\langle U(\cdot), V(\cdot)\rangle := \int U(n)V(n)d\Omega(n)$, where the integration is over the sphere with $\int d\Omega(n) = 1$. Then, because $\mathcal{V}_{l+1} \subset \mathcal{V}_l$, any field $T_l(n) \in \mathcal{V}_l$ can be decomposed as

$$T_l(n) = T_{l+1}(n) + D_l(n), \tag{3}$$

where $T_{l+1}(n) \in \mathcal{V}_{l+1}$ is the projection of $T_l(n)$ onto $\mathcal{V}_{l+1}$ and $D_l(n) \in \mathcal{W}_l := \mathcal{V}_l \ominus \mathcal{V}_{l+1}$ is orthogonal to $\mathcal{V}_{l+1}$. Because $\mathcal{V}_{l+1}$ is obtained by removing the SBFs on $\mathcal{M}_l$ from $\mathcal{V}_l$, the orthogonal complement $\mathcal{W}_l$ can be interpreted as containing the local information near $\mathcal{M}_l$ that can not be explained by the space $\mathcal{V}_{l+1}$ that contains the global trend extrapolated from the sparser network $\mathcal{N}_{l+1}$. Therefore, $T_{l+1}(n)$ is called the global component of scale $l+1$ and $D_l(n)$ is called the local component of scale l.

Because $\mathcal{V}_1 \supset \mathcal{V}_2 \supset \cdots \supset \mathcal{V}_L$ and

$$\mathcal{V}_l = \mathcal{V}_{l+1} \oplus \mathcal{W}_l \qquad (l = 1, \ldots, L-1), \tag{4}$$

it follows from (3) that $T_1(n)$ in (2) can be decomposed as

$$T_1(n) = T_L(n) + \sum_{l=1}^{L-1} D_l(n), \tag{5}$$

where $T_L(n) \in \mathcal{V}_L$ and $D_l(n) \in \mathcal{W}_l$. Note that the $D_l(n)$ are orthogonal to each other as well as to $T_L(n)$. More details of multiresolution analysis based on SBF representation are discussed by Li (1999).

2.2 MULTISCALE SPHERICAL WAVELETS

The orthogonal complement $\mathcal{W}_l$ can be characterized by spherical wavelets. Li (1999) showed that with certain filters $e_l(i,j)$, which depend only on the SBFs and the subnetworks, the spherical wavelets defined by

$$W_{lj}(n) := G_l(n \cdot n_{lj}) - \sum_{k=l+1}^{L} \sum_{i=1}^{M_k} e_k(i,j) G_k(n \cdot n_{ki}) \tag{6}$$

completely determine $\mathcal{W}_l$ such that

$$\mathcal{W}_l = \operatorname{span}\{W_{lj}(n) : j = 1, \ldots, M_l\}.$$

As an important feature for spatial adaptivity, it can be shown (Li, 1999) that under suitable conditions $W_{lj}(n)$ is localized near n_j and the degree of localization is proportional to the bandwidth of $G_l(n \cdot n_{lj})$.

Because $D_l(n)$ can be expressed as a linear combination of the SWs that characterize $\mathcal{W}_l$, $T_1(n)$ in (5) has an equivalent SW representation

$$T_1(n) = \sum_{j=1}^{M_L} \beta_{Lj} G_L(n \cdot n_{Lj}) + \sum_{l=1}^{L-1} \sum_{j=1}^{M_l} \gamma_{lj} W_{lj}(n). \tag{7}$$

In commonly-used wavelet terminology, the β_{Lj} in (7) can be considered as smooth coefficients and the γ_{lj} as detail coefficients. These coefficients can be computed from the β_{1j} in (2) by a recursive algorithm (Li, 1999).

In the SW representation (7), a spherical field is decomposed into multiscale components which are orthogonal with respect to the scales. This multiscale decomposition makes the SW representation a potentially useful tool in many applications such as compressing spherical data, detecting local anomalies, and multiscale dynamic modelling.

3 Networks and Bandwidth of SBFs

When we investigate the potential of the multi-scale SW representation for a climate field, the appropriate choice of networks $\mathcal{N}_l$ and SBFs $G_l(x)$ are required to describe the field effectively.

3.1 THE DESIGN OF NETWORKS

In this section, we suggest some schemes to choose the nested networks $\mathcal{N}_l$ systematically for performing the bottom-up design approach. Our network design depends only on the location of data and the type of grid which is predetermined without considering geophysical information.

Before explaining the steps of network design, we introduce two types of grid: a standard grid and Göttelmann's grid. Each grid type has a regular and a reduced grid. The reduced grid is designed to overcome the regular grid problem of a strong concentration of points near the poles. But since the network is selected by the relation of data-observing sites and a grid type, as will be mentioned later, we can not assure that the reduced grid produces networks that can better represent spherical fields. Let us define $[0, 2\pi]$ as the range of longitude and $[0, \pi]$ as the range of latitude to obtain grid points. Then, by simple transformation, all grid points can be located on $[-180°, 180°]$ as longitude and $[-90°, 90°]$ as latitude.

3.1.1 *Standard Regular Grid*

For $l \in \mathbb{N}$, we define the index set

$$K_l^s := \{(i,k) : k = 0, 1, \ldots, 2^l; i = 0, 1, \ldots, 2^{l+1}\}, \tag{8}$$

where l is the index of level, i is the index of grid point of longitude and k denotes the index of grid point of latitude. The regular grid induced by the index set in (8) is defined

$$\mathcal{T}_l^s := \{(\phi_{i,l}, \theta_{k,l}); (i,k) \in K_l^s\}, \tag{9}$$

where grid points of longitude, $\phi_{i,l} = i\frac{\pi}{2^l}$ and grid points of latitude, $\theta_{k,l} = k\frac{\pi}{2^l}$. Obviously, the sequence $\{\mathcal{T}_l^s\}_{l\in\mathbb{N}}$ of grids is hierarchical: $\mathcal{T}_l^s \subseteq \mathcal{T}_{l+1}^s$ for all $l \in \mathbb{N}$. For the simplest example, let $l = 1$, the grid $\mathcal{T}_1^s$ is

$$\mathcal{T}_1^s := \{(\phi_{i,1}, \theta_{k,1}); (i,k) \in K_1^s\},$$

where $\phi_{i,1} = \{0, \frac{\pi}{2}, \pi, \frac{3\pi}{2}, 2\pi\}$, $\theta_{k,1} = \{0, \frac{\pi}{2}, \pi\}$ and $K^s_1 = \{(i,k) : i = 0,1,2,3,4; k = 0,1,2\}$.

3.1.2 *Standard Reduced Grid*
The index set $K^{s,r}_l$ for reduced grid is

$$K^{s,r}_l := \{(i,k) : k = 0,1,\dots,2^l; i = 0,1,\dots,\frac{2^{l+1} - 2^{r^s_{k,l}}}{2^{r_{k,l}}} + 1\} \subseteq K^s_l, \tag{10}$$

where the control parameter for dropping grid points according to latitude level, $r^s_{k,l}$ is defined to be

$$r^s_{k,l} = \begin{cases} 0, & \frac{\pi}{4} \leq \theta_{k,l} \leq \frac{3\pi}{4} \\ \lfloor l - \log_2(\pi k) \rfloor, & 0 < \theta_{k,l} < \frac{\pi}{4} \\ \lfloor l - \log_2(\pi(2^l - k)) \rfloor, & \frac{3\pi}{4} < \theta_{k,l} < \pi \\ l - 1, & \theta_{k,l} = 0, \pi \end{cases} \tag{11}$$

Here $\lfloor x \rfloor$ denotes the largest integer less or equal to x. The reduced grid is given by

$$\mathcal{T}^{s,r}_l := \{(\phi_{i,l}, \theta_{k,l}); (i,k) \in K^{s,r}_l\} \subseteq \mathcal{T}^s_l, \tag{12}$$

where grid points of longitude, $\phi_{i,l} = i2^{r^s_{k,l}} \frac{\pi}{2^l}$ and grid points of latitude, $\theta_{k,l} = k\frac{\pi}{2^l}$.

3.1.3 *Göttelmann's Regular Grid*
The index set K^G_l is defined as

$$K^G_l := \{(i,k) : k = 0,1,\dots,2^{l+1}; i = 0,1,\dots,2^{l+2}\}. \tag{13}$$

By setting grid points of longitude and grid points of latitude

$$\phi_{i,l} = i\frac{\pi}{2^{l+1}} \quad \text{and} \quad \theta_{k,l} = k\frac{\pi}{2^{l+1}},$$

we obtain the grid $\mathcal{T}^G_l := \{(\phi_{i,l}, \theta_{k,l}); (i,k) \in K^G_l\}$. The grid $\mathcal{T}^G_l$ is the same as $\mathcal{T}^s_{l+1}$

3.1.4 *Göttelmann's Reduced Grid*
Let grid points of longitude, $\phi_{i,l} = i2^{r^G_{k,l}} \frac{\pi}{2^{l+1}}$ and grid points of latitude, $\theta_{k,l} = k\frac{\pi}{2^{l+1}}$. The control parameter, $r^G_{k,l}$ is defined to be

$$r^G_{k,l} = \begin{cases} 0, & \frac{\pi}{4} \leq \theta_{k,l} \leq \frac{3\pi}{4} \\ \lfloor l + 1 - \log_2(\pi k) \rfloor, & 0 < \theta_{k,l} < \frac{\pi}{4} \\ \lfloor l + 1 - \log_2(\pi(2^{l+1} - k)) \rfloor, & \frac{3\pi}{4} < \theta_{k,l} < \pi \\ l - 1, & \theta_{k,l} = 0, \pi \end{cases} \tag{14}$$

Thus, the reduced grid is given by

$$K^{G,r}_l \ := \ \{(i,k) : k = 0,1,\dots,2^{l+1}; i = 0,1,\dots,\frac{2^{l+2} - 2^{r^G_{k,l}}}{2^{r^G_{k,l}}} + 1\} \subseteq K^G_l.$$

$$\mathcal{T}_l^{G,r} \ := \ \{(\phi_{i,l}, \theta_{k,l}); (i,k) \in K_l^{G,r}\} \subseteq \mathcal{T}_l^G. \tag{15}$$

The grid $\mathcal{T}_l^{G,r}$ is similar to the grid $\mathcal{T}_{l+1}^{s,r}$. But the grid points near the two poles are different from $\mathcal{T}_{l+1}^{s,r}$.

Figure 2 illustrates the grid points obtained from Göttelmann's grid.

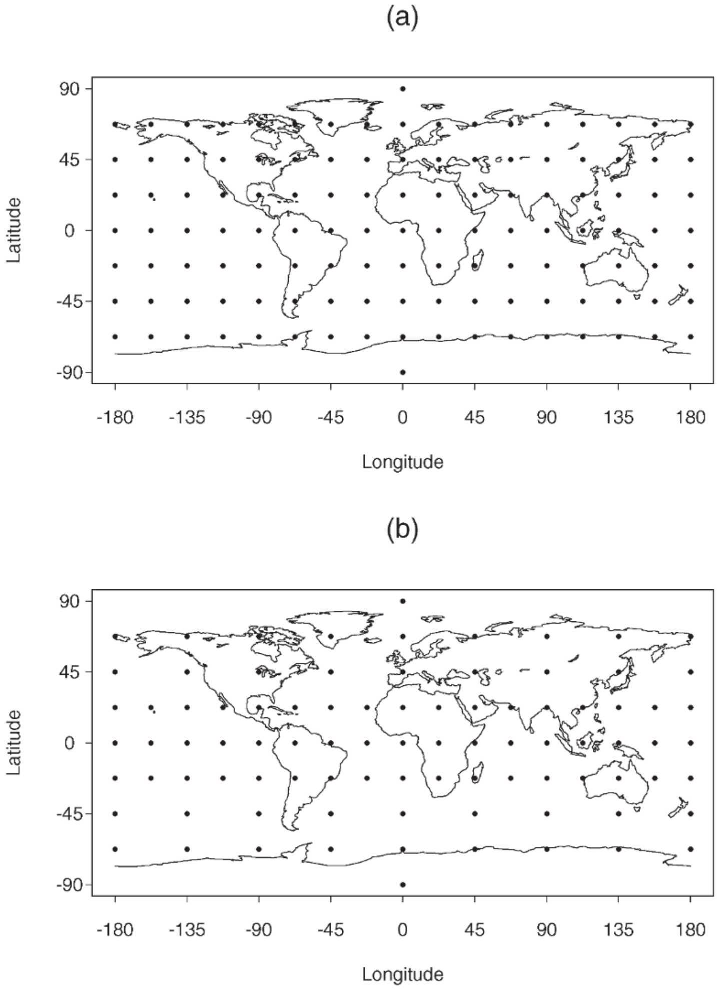


Figure 2. Grid points of Göttelmann's grid for $l = 2$. (a) Regular grid points; (b) Reduced grid points.

Our network is designed systematically based on the location of data and the type of grid. As the resulting network, we expect that in each level (resolution) $\mathcal{N}_L$ or $\mathcal{M}_l$, stations are distributed over the sphere as uniformly as possible, and stations between two levels are not too close so that we can apply SBFs with different bandwidths to these stations. Here are the steps of network design.

1. Obtain the center points c_i of each grid box from the grid $\mathcal{T}_1$ which has the smallest number of grid points. Form a territory (circle) D_i around a given center point. Thus the number of territories should be equal to the number of center points. Then compute the geodesic distance from locations of data $d_j \in D_i$ within a territory to the center point of the territory, $\{\arccos(\theta(c_i \cdot d_j))\}$ for $i = 1, 2, \ldots, N_L$ and find the location of data which has the minimum distance to the center point. A network with sparsely located stations, $\mathcal{N}_L$ is made up from these locations.

2. From the next grid $\mathcal{T}_2$, compute the center point c_i of each grid box. As with step 1, draw a territory (circle) D_i around a given center point and find the closest station to the center point within its territory, D_i, for all i. Thus, the set $\mathcal{M}^*_{L-1}$ is obtained from the selected stations. Then compare these stations with stations in $\mathcal{N}_L$. That is, compute the geodesic distance from the stations in $\mathcal{M}^*_{L-1}$ to the stations in $\mathcal{N}_L$. If some distance between two locations is closer than a criterion, the location is deleted from $\mathcal{M}^*_{L-1}$. After the comparison procedure, the network $\mathcal{M}_{L-1}$ is formed. Note that $\mathcal{N}_{L-1} = \mathcal{N}_L \cup \mathcal{M}_{L-1}$.

3. Repeat step 2 for remaining data and grid points from $\mathcal{T}_l$ for $l = 2, 3, \ldots$ until the longitude of grid box from $\mathcal{T}_l$ becomes 5 degrees. Finally, $\mathcal{N}_1 = \mathcal{N}_L \cup \mathcal{M}_{L-1} \cup \cdots \cup \mathcal{M}_1$ is obtained.

For performing the above steps, we need to discuss the territory and criterion for each step. The radius of territory should not be too large to reduce the possibilities which select stations located far from the center point of each grid box. Because when there are no stations in a given grid box, and the radius of territory is too large, a station in another grid box can be selected as closest stations to the center point of this grid box.

Figure 3 shows a set of networks with 6 levels for the stations which recorded the surface air temperature of 1973. These networks are obtained from the standard regular grid.

3.2 THE CHOICE OF BANDWIDTHS

We now discuss the scale parameter (bandwidth) of SBFs to be employed for each network $\mathcal{N}_L, \mathcal{M}_{L-1}, \ldots, \mathcal{M}_1$.

First of all, let us look at the scale parameter of 1-d wavelets defined on a real line in order to preview the bandwidth of spherical wavelets and to understand the procedure of the choice of bandwidths to be explained later. From the definition of 1-d wavelets

$$\psi_{j,k}(t) = 2^{-j/2}\psi\left(\frac{t - 2^j k}{2^j}\right),$$

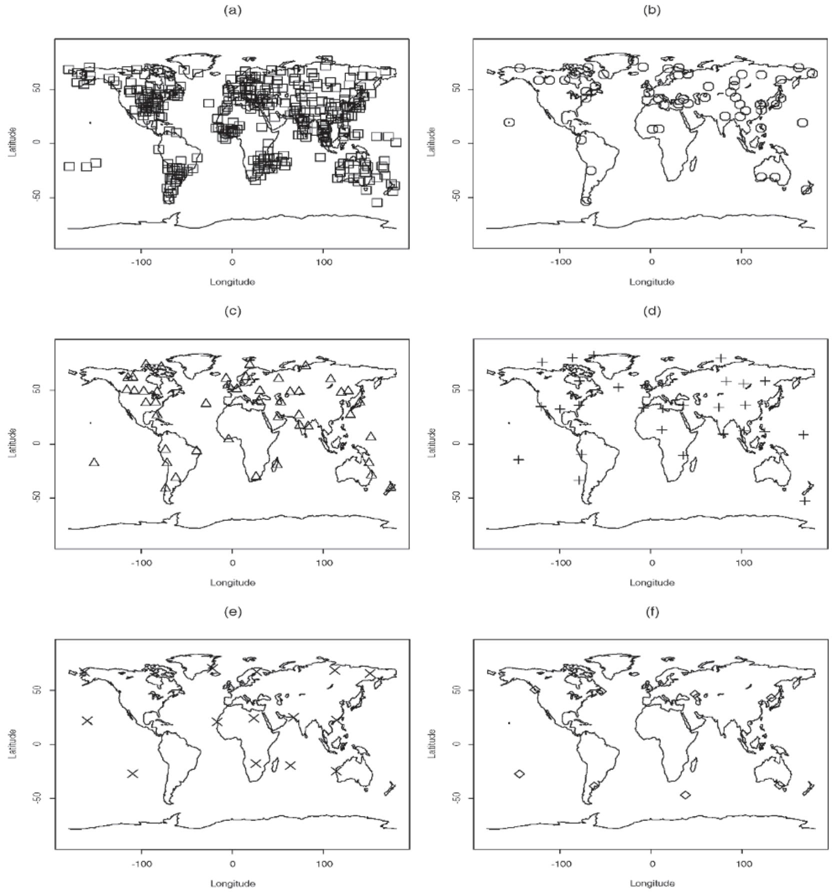


Figure 3. Networks for 1973 surface air temperature stations. (a) Network $\mathcal{M}_1$; (b) Network $\mathcal{M}_2$; (c) Network $\mathcal{M}_3$; (d) Network $\mathcal{M}_4$; (e) Network $\mathcal{M}_5$; (f) Network $\mathcal{N}_6$.

we know that the scale parameter j is decided systematically. That is, as the scale parameter j increases 1 unit every time, the length of support of the wavelet increases twice as much as before. Thus there is a relationship between the scale parameter and the length of the wavelet.

Similarly, we suggest that the bandwidths η_l of SBFs be chosen such that

$$\eta_{L-l} = e^{-\rho_l}, \quad l = 1, 2, \ldots, L-1, \tag{16}$$

where $\rho_l = \frac{\rho^*}{2^l}$. The ρ^* can be obtained from the bandwidth of the smallest network level L, η_L by $\rho^* = -\log \eta_L$. As mentioned before, the networks are related to a grid. As the level l decreases, a grid box related to the level decreases one fourth in size (both intervals of latitude and longitude decrease by a half). Thus, the area covered by a SBF decreases as the level l decreases. That is the reason we use the (16) bandwidths η_l. Hence, if the bandwidth η_L of the sparsest level L is decided, all bandwidths can be obtained systematically.

Now let us discuss how we can get the bandwidth η_L. From simple geometry, the surface area covered by surface mass distribution with variance σ^2 over unit sphere Ω is $2\pi(1-\sqrt{1-\sigma^2})$. Since the total surface area of the unit sphere is 4π and the variance of SBF or spherical wavelet from Poisson kernel is $\sigma^2 = \left(\frac{1-\eta^2}{1+\eta^2}\right)^2$, the surface area covered is $2\pi\left(1-\sqrt{1-\left(\frac{1-\eta^2}{1+\eta^2}\right)^2}\right)$. Note that as the bandwidth of SBF is close to 0, the surface area covered becomes 2π. In this case, we need two stations to cover the whole sphere Ω. In another extreme case, as the η is going to 1, the surface area covered is close to 0. To cover the whole sphere Ω, we need ∞ stations. Under the assumption that the stations are distributed equally over the sphere, it can be easily known how many stations are needed in order to cover the whole sphere with fixed η and how large the bandwidth of SBF is needed to cover the whole sphere when the number of stations are fixed from the following

$$\#\text{ of stations} = n = \frac{2}{1-\sqrt{1-\left(\frac{1-\eta^2}{1+\eta^2}\right)^2}} \quad \text{and} \quad \eta = \sqrt{\frac{1-a_n}{1+a_n}},$$

where $a_n = \sqrt{1-\left(1-\frac{2}{n}\right)}$. Stations in the sparsest network $\mathcal{N}_L$ based on any grid scheme are almost distributed equally. Thus, from the above equations, we can decide the bandwidth η_L. For example, we decide SBFs with 6 different bandwidths to be employed at networks with six levels in Figure 3. The result is that $\eta_1 = 0.9596$ for $\mathcal{M}_1$, $\eta_2 = 0.9209$ for $\mathcal{M}_2$, $\eta_3 = 0.8482$ for $\mathcal{M}_3$, $\eta_4 = 0.7194$ for $\mathcal{M}_4$, $\eta_5 = 0.5176$ for $\mathcal{M}_5$ and $\eta_6 = 0.2679$ for $\mathcal{N}_6$.

4 Experimental Study

4.1 THE CHOICE OF COEFFICIENTS

The coefficients $\boldsymbol{\beta}_1 = \text{vec}\{\beta_{1j}\}$ of (2) can be obtained in many ways. The simplest example is observations $\boldsymbol{t} = \text{vec}\{T(n_j)\}$. Once the $T_1(n)$ is obtained, all global fields $T_l(n)$ for $l = 2,\ldots,L$ and detail fields $D_l(n)$ for $l = 1,2,\ldots,L-1$ is decided by the multiresolution analysis. Thus, whether or not the $T_1(n)$ has a good performance is very important for a good representation.

We now discuss the interpolation methods for obtaining coefficients $\boldsymbol{\beta}_1$. In (2), because the interpolation matrix $\boldsymbol{G}_1 = [G_{l'}(n_i \cdot n_{l'j})]_{i,j}$ for $l^{'} = 1,\ldots,L$ is invertible, we can get $\tilde{\boldsymbol{\beta}}_1 = \text{vec}\{\tilde{\beta}_{1j}\} = (\boldsymbol{G}_1^T\boldsymbol{G}_1)^{-1}\boldsymbol{G}_1^T\boldsymbol{t}$ by the least squares method. The solution $\tilde{\boldsymbol{\beta}}_1$ gets values through all data points at the observation sites. Thus $T_1(n)$ is an interpolation function. As another example, let us consider the interpolation method by penalized least squares. The penalized least squares solution is $\tilde{\boldsymbol{\beta}}_1^p = \text{vec}\{\tilde{\beta}_{1j}^p\} = (\boldsymbol{G}_1^T\boldsymbol{G}_1 + \lambda\boldsymbol{Q})^{-1}\boldsymbol{G}_1^T\boldsymbol{t}$. The λ should be selected appropriately. Note that the penalized least squares becomes the ridge regression method by setting $\boldsymbol{Q} = \boldsymbol{I}$.

4.2 MULTISCALE SW REPRESENTATION FOR A TEMPERATURE FIELD

The multiscale SW representation for the 1973 winter global temperatures is displayed in Figure 4. Figure 5 shows spherical smoothing spline estimate. As expected, the SW representation captures a global pattern of Northern Hemisphere winter mean temperatures across the world with local activities in some regions (e.g, Siberia as the coldest spot, North-West region of Australia as the hottest). The spine estimate, on the other hand, trends to be uniformly smooth across the globe. Notice that for multiscale SW representation, the standard regular grid with six levels is adapted for networks and coefficients $\boldsymbol{\beta}_1$ is obtained from the LS interpolation method. For spherical smoothing spline method, cross-validation has been used for selecting smoothing parameter.

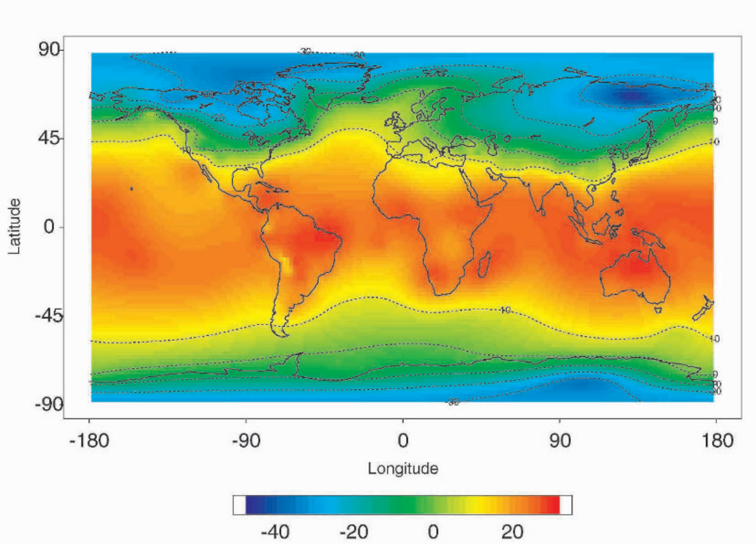


Figure 4. Multiscale SW representation for 1973 winter global temperatures.

5 Conclusions

To apply the multiscale SW representation effectively, the appropriate choice of bandwidth of SBFs and networks is necessary. In this paper, we have proposed methods to choose the nested networks $\mathcal{N}_l$ and the scale parameter (bandwidth)

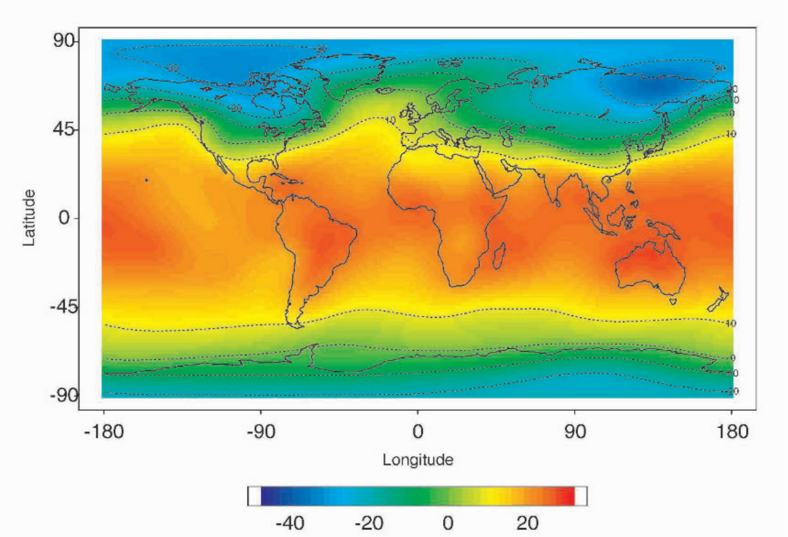


Figure 5. Smoothing spline representation for 1973 winter global temperatures.

of SBFs to be employed for each network systematically. Our approach depends only on the location of data and the type of grid which is predetermined without considering geophysical information. Bandwidths of SBFs are decided on by a systematic method to make the best use of the advantage of wavelets which has a fast algorithm. Based on the above practical issues for implementation of multiscale SW, we have investigated the potential of the multiscale SW representation for the surface air temperature data.

In the experimental study for global surface air temperatures, we have shown that multiscale SW estimators are very powerful for detecting local activities as well as extracting global trends of temperature fields that cannot be easily detected by the traditional spherical smoothing spline method.

As related future researches, a network modified by geophysical information should produce a more effective multiscale SW representation. We can expect that the bandwidth with more adaptable method such as generalized cross-validation (GCV) or cross-validation (CV) improves the performance of the multiscale SW representation.

Acknowledgements

This work was supported in part by the grant ATM-0327936 of U.S. National Science Foundation and the Natural Sciences and Engineering Council of Canada.

This page intentionally blank

References

Cressie, N. A. C., *Statistics for Spatial Data (Revised Edition)*, Wiley-Interscience, 1993.

Freeden, W., Gervens, M. and Schreiner, M., *Constructive Approximation on the Sphere with Applications to Geomathematics*, Oxford University Press, 1998.

Jones, P.D., Raper, S.C.B., Cherry, B.S.G., Goodess, C.M., Wigley, T.M.L, Santer, B., Kelly, P.M., Bradley, R.S. and Diaz, H.F., *An updated global grid point surface air temperature anomaly data set: 1851-1988*, Environmental Sciences Division Publication No. 3520, U.S. Department of Energy, 1991.

Li, T.H., *Multiscale representation and analysis of spherical data by spherical wavelets*, SIAM Journal of Scientific Computing, vol. 21, 1999, p. 924-953.

Li, T.H., *Multiscale wavelet analysis of scattered spherical data: design and estimation*, Environmetrics, vol. 12, 2001, p. 179-202.

Narcowich, F.J. and Ward, J.D., *Nonstationary wavelets on the m- sphere for scattered data*, Applied and Computational Harmonic Analysis, vol. 3, 1996, p. 324-336.

MULTIVARIATE GEOSTATISTICAL MAPPING OF ATMOSPHERIC DEPOSITION IN FRANCE

OLIVIER JAQUET[1], LUC CROISÉ[2], ERWIN ULRICH[2] AND PIERRE DUPLAT[2]
[1]*Colenco Power Engineering Ltd, Taefernstrasse 26, 5405 Baden, Switzerland*
[2]*Office National des Forêts, Département Recherche et Développement, Boulevard de Constance, 77300 Fontainebleau, France*

Abstract. This study presents a multivariate geostatistical approach for the mapping of atmospheric deposition at the scale of the entire French territory, including precipitation as an auxiliary variable. By applying cokriging, deposition maps were produced with the corresponding uncertainty and with an improved level of detail in comparison to previous studies.

1 Introduction

At the beginning of the eighties, central European forests showed unusual crown deterioration described as a "new type of forest decline". At that time, suggestions linked this deterioration mainly to air pollution. The negative impact of acid deposition on the functioning of several ecosystems constitutes a reality (Johnson and Lindberg, 1992; Draaijers et al., 1997). Due to the limited capacity of forest ecosystems to balance acidic atmospheric inputs, in the nineties, the concept of "critical loads" was raised as a tool to elaborate protocols for pollutant emissions reductions in Europe. In the last decade, in France (Party et al., 2001) and in other countries (Posch et al., 2001), important efforts were performed for the characterisation and mapping of the critical loads for acidity. In order to obtain such information, it was essential to develop a simple and reliable methodology to produce accurate maps of the atmospheric deposition in order to identify the sites for which critical loads are in exceedance.

Within the framework of the methodology developed by the ONF (Office National des Forêts), the monitoring of ecological parameters is achieved via the RENECOFOR network (REseau National de suivi à long terme des ECOsystèmes FORestiers). Chemical analyses of atmospheric deposition have been conducted since 1993 using the sub-network CATAENAT (Charge Acide Totale d'origine Atmosphérique sur les Ecosystèmes NAturels Terrestres). The objectives of this measurement campaign are the monitoring of: (1) potential evolutions of atmospheric deposition in time and (2) the spatial distributions of this deposition for France in relation to critical loads for acidity. Using six years worth of measurements from the CATAENAT network has allowed us to develop deterministic models for deposition (Croisé et al., 2002) with only a few

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 833–841.

explanatory variables (precipitation, altitude and period of the year). The application of this method has resulted in the first reliable deposition maps for France.

The use of geostatistical methods was motivated by accounting for spatial correlation and by producing deposition maps for all of France and not only for discrete locations as in previous attempts (Croisé et al., 2002). Furthermore, the use of geostatistical methods constitutes a valuable novel alternative for mapping deposition within a probabilistic framework. The application and evaluation of these methods for the spatial characterisation of atmospheric deposition is the subject of this study.

2 Multivariate geostatistical mapping

Mapping of a regionalised variable (Re.V.) such as atmospheric deposition requires the interpolation of the data at block centres (i.e., surface centres in 2D) of a regular grid. Since, the Re.V. is considered to be a realisation of a random function, the interpolation method applies an estimator obtained from a linear combination of the different random variables (R.V.) representing the data:

$$Z^*_{i_0}(s_0) = \sum_{\mathrm{i=1}}^{\mathrm{N}} \sum_{\alpha=1}^{n_i} \lambda^i_\alpha Z_i(x_\alpha)$$

where:

$Z^*_{i_0}(s_0)$: estimator of the main R.V., i_0, for the bloc s_0

λ^i_α : weights (unknowns)

$Z_i(x_\alpha)$: R.V. representing main and auxiliary Re.V.

At the data level, n_i represents the number of values for a given Re.V. of index i. N corresponds to the number of Re.V. The weights are obtained by solving the linear equations of the ordinary cokriging system (Wackernagel, 1995):

$$\sum_{j=1}^{N} \sum_{\beta=1}^{n_j} \lambda^j_\beta \gamma_{ij}(x_\alpha - x_\beta) + \mu_i = \bar{\gamma}_{ii_0}(x_\alpha, s_0)$$

$$\sum_{\beta=1}^{n_i} \lambda_\beta = \delta_{ii_0} = \begin{cases} 1 & if \quad i = i_0 \\ 0 & if \quad i \neq i_0 \end{cases}$$

where:

$\gamma_{ij}(x_\alpha - x_\beta)$: variograms and cross variograms

$\bar{\gamma}_{ii_0}$: mean variograms and cross variograms

μ_i : Lagrange multipliers.

With the advantage of providing the cokriging variance which quantifies the spatial uncertainty associated with the interpolation:

$$\sigma^2_{ck}(s_0) = \sum_{i=1}^{N} \sum_{\alpha=1}^{n_i} \lambda^j_\alpha \bar{\gamma}_{ii_0}(x_\alpha, s_0) + \mu_{i_0} - \bar{\gamma}_{ii_0}(s_0, s_0)$$

where:

$\sigma^2_{ck}(s_0)$: cokriging variance for the bloc s_0 .

Prior to the cokriging stage, the experimental variograms and cross variograms are estimated from the data. Then, they are fitted semi-automatically by least squares using authorised functions within the framework of the linear model of coregionalisation which requires particular algebraic conditions for its parameters to be satisfied (Chilès and Delfiner, 1999). When more than one auxiliary variable is considered, an iterative algorithm by Goulard (1989), implemented in the Isatis (2002) software, is applied for model fitting.

The model performance regarding the contribution of auxiliary variable(s) is evaluated by cross validation (Wackernagel, 1995) using the following criteria:

$$MSE = \frac{1}{\mathrm{n}} \sum_{\alpha=1}^{\mathrm{n}_{\mathrm{i}_0}} [Z^*_{i_0}(x_{(\alpha)}) - Z_{i_0}(x_\alpha)]^2$$

$$VR = \frac{1}{\mathrm{n}} \sum_{\alpha=1}^{\mathrm{n}_{\mathrm{i}_0}} \frac{[Z^*_{i_0}(x_{(\alpha)}) - Z_{i_0}(x_\alpha)]^2}{\sigma^2_{\mathrm{ck}(\alpha)}}$$

Where:

MSE : mean square error

$Z^*_{i_0}(x_{(\alpha)})$: estimator of the main R.V. at measurement site x_α

$Z_{i_0}(x_\alpha)$: data value at measurement site x_α

VR : variance ratio

$\sigma^2_{ck(\alpha)}$: cokriging variance for measurement site x_α .

3 Mapping of atmospheric deposition

3.1 DATA

The ONF (Office National des Forêts) is responsible for the monitoring of ecological parameters using the RENECOFOR network (REseau National de suivi à long terme des ECOsystèmes FORestiers). Chemical analyses of atmospheric deposition have been conducted since 1993 using the 27 stations of the sub-network CATAENAT (Charge Acide Totale d'origine Atmosphérique sur les Ecosystèmes NAturels Terrestres) all throughout France (Figure 1). All sites are located in forested areas, out in the open, and rather far away from point emission sources. The distribution of the measuring sites provides for coverage of all forest regions in France. The altitude of the sites ranges between 5 and 1400 m a.s.l.. Representative weighted samples of deposition were obtained by mixing weekly samples over a period of four weeks for each site. For these samples, the major anions (S-SO_4, N-NO_3 and Cl) and cations (N-NH_4, Ca, Mg, K and Na) were analysed as well as the protons (H).

In parallel to deposition sampling, precipitation was measured at each of the 27 sites. In addition, daily precipitation data were also available for 2561 stations of the meteorological network of Météo-France (Figure 1).

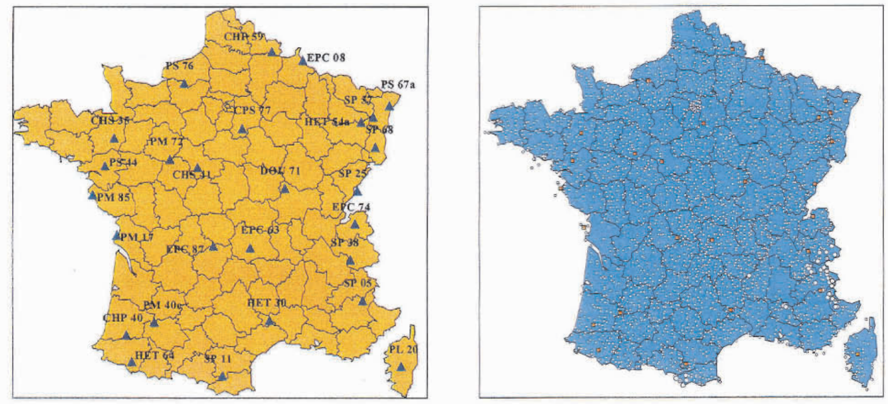


Figure 1. Geographical distribution of the CATAENAT measuring sites for deposition and precipitation (left) and of the Météo-France precipitation stations (right).

The data used in this study were recorded over a six year period between 1993 and 1998.

3.2 REGIONALISED VARIABLES

For the data sampled at the 27 sites of the CATAENAT network, 11 regionalised variables were considered for the six year period: the annual mean deposition for nine ions (S-SO_4, N-NO_3, Cl, N-NH4, Ca, Mg, K, Na and H), the annual mean precipitation (P) and the altitude of the measurement stations (Z). In addition, annual mean precipitation and altitude (below 1400 m) data taken from the Météo-France network were available for 2561 stations.

The deposition variables constitute the main variables; precipitation and altitude are taken as auxiliary variables. The main variables are in partial heterotopy with respect to all sampling locations; i.e., these variables are only available at 27 locations as opposed to precipitation and altitude which were measured at all 2588 stations (Table 1).

3.3 VARIOGRAMS

The statistics of the 11 main and auxiliary variables are given in Table 1. For the 11 variables, 11 experimental variograms are calculated as well as 19 experimental cross variograms corresponding to the selected pairs; i.e., one main variable (deposition) associated with one of the two auxiliary variables (precipitation or altitude).

Variable	Minimum	Maximum	Mean	Std. dev.	Variance	Nb.*	Network
1 : S-SO4	*3.7 kg/ha/yr*	*15.9 kg/ha/yr*	*6.7 kg/ha/yr*	*2.4 kg/ha/yr*	*5.8 (kg/ha/yr)*^{2}	*27*	CATAENAT
2 : N-NO3	*1.8 kg/ha/yr*	*6.9 kg/ha/yr*	*3.5 kg/ha/yr*	*1.2 kg/ha/yr*	*1.4 (kg/ha/yr)*^{2}	*27*	CATAENAT
3 : Cl	*3 kg/ha/yr*	*84 kg/ha/yr*	*21 kg/ha/yr*	*22 kg/ha/yr*	***496*** *(kg/ha/yr)*^{2}	*27*	CATAENAT
4 : N-NH4	*1.9 kg/ha/yr*	*10.9 kg/ha/yr*	*5.0 kg/ha/yr*	*2.1 kg/ha/yr*	*4.4 (kg/ha/yr)*^{2}	*27*	CATAENAT
5 : Ca	*2.6 kg/ha/yr*	*15.1 kg/ha/yr*	*6.0 kg/ha/yr*	*3.3 kg/ha/yr*	*11.0 (kg/ha/yr)*^{2}	*27*	CATAENAT
6 : Mg	*0.2 kg/ha/yr*	*6.6 kg/ha/yr*	*1.7 kg/ha/yr*	*1.7 kg/ha/yr*	*3.1 (kg/ha/yr)*^{2}	*27*	CATAENAT
7 : K	*0.6 kg/ha/yr*	*3.8 kg/ha/yr*	*1.6 kg/ha/yr*	*0.7 kg/ha/yr*	*0.4 (kg/ha/yr)*^{2}	*27*	CATAENAT
8 : Na	*2 kg/ha/yr*	*48 kg/ha/yr*	*12 kg/ha/yr*	*13 kg/ha/yr*	***165*** *(kg/ha/yr)*^{2}	*27*	CATAENAT
9 : H	*21 g/ha/yr*	*353 g/ha/yr*	*115 g/ha/yr*	*67 g/ha/yr*	***4462*** *(g/ha/yr)*^{2}	*27*	CATAENAT
10 : P	*452.0 mm*	*2765.9 mm*	*959.4 mm*	*289.9 mm*	*84048.8 mm*^{2}	*2588*	CATAENAT + Météo-France
11 : Z	*1 m*	*1400 m*	*309 m*	*300.8 m*	*90507 m*^{2}	*2588*	CATAENAT + Météo-France

*Nb.: number of measurement sites/stations.

Table 1. Statistics for the main and auxiliary variables

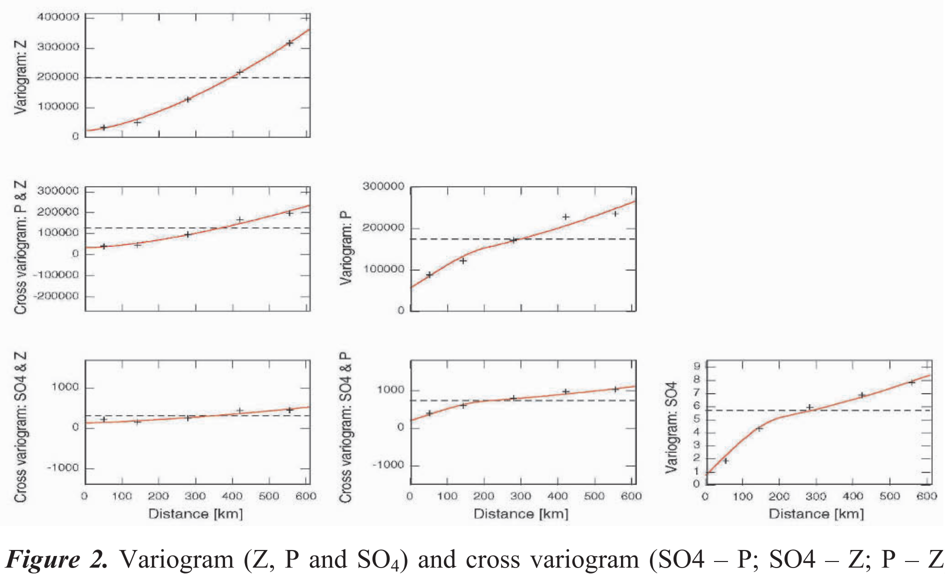


Figure 2. Variogram (Z, P and SO$_4$) and cross variogram (SO4 – P; SO4 – Z; P – Z) fitted to the experimental variograms and cross variograms (crosses) using the linear model of coregionalisation. The dashed horizontal line corresponds to the variance/covariance of the data.

The observed shapes of the experimental variograms indicate correlated and cross-correlated spatial behaviours, respectively, for the variables as well as for the variable pairs studied (Figure 2); i.e., with increasing distance, the variograms and cross variograms display a growth without apparent stabilisation at the working scale of about 600 km. Such behaviours were observed for all deposition variables. Despite the small number of stations, the correlated spatial behaviours shown by the experimental variograms and cross variograms were consistent with deposition patterns (Croisé et al., 2002). These results allow the application of geostatistical methods for the mapping of atmospheric deposition.

The fitting of all variograms and cross variograms was carried out in the framework of the linear model of coregionalisation (cf. section 2) using a nested variogram composed of nugget effect, and spherical and power components. Therefore, three models are available for each deposition variable: a univariate model for deposition and two multivariate models with auxiliary variables (deposition + precipitation; deposition + precipitation + altitude).

3.4. CROSS VALIDATION

The best model should present the smallest mean square error and a variance ratio as close as possible to unity (cf. section 2). Cross validation was carried out for the nine univariate models as well as for the 18 multivariate models with auxiliary variables (Table 2).

Main variable[1]	**Auxiliary Variable(s)**	**MSE** [kg/ha/year]^2	**VR** [-]	**Remark**
S-SO4	-	*6.3*	*1.3*	-
S-SO4	P	*3.5*	*1.0*	-
S-SO4	P + Z	*3.6*	*1.3*	-
N-NO3	-	*1.6*	*1.7*	-
N-NO3	P	*0.7*	*1.0*	-
N-NO3	P + Z	*0.7*	*1.2*	-
Cl	-	*356*	*1*	-
Cl	P	*341*	*1*	-
Cl	P + Z	*304*	*1*	*kriged values < 0*
N-NH4	-	*3.3*	*1.2*	-
N-NH4	P	*2.7*	*1.3*	-
N-NH4	P + Z	*2.7*	*1.3*	-
Ca	-	*7.5*	*1.3*	-
Ca	P	*7.6*	*1.6*	-
Ca	P + Z	*7.2*	*1.5*	-
Mg	-	*2.0*	*1.3*	-
Mg	P	*1.9*	*1.4*	-
Mg	P + Z	*1.8*	*1.5*	*kriged values < 0*
K	-	*0.5*	*1.6*	-
K	P	*0.4*	*1.5*	-
K	P + Z	*0.4*	*1.5*	-
Na	-	*122*	*1*	-
Na	P	*101*	*1*	-
Na	P + Z	*97*	*1*	*kriged values < 0*
H	-	*5311*[2]	2	-
H	P	*3033*[2]	*1*	-
H	P + Z	*3033*[2]	*1*	-

[1] in grey, the selected models; [2] units are [g/ha/year]^2

Table 2. Cross validation results.

The results show that in general the multivariate model (deposition + precipitation) delivers the best cross validation scores. Some models present better scores in the presence of two auxiliary variables. However, when applying these models for mapping, some of the cokriged values, in zones with low altitude and poor atmospheric deposition, can become negative. These results, although being statistically acceptable, are physically erroneous. Consequently, these multivariate models were disregarded.

3.5 COKRIGING OF ATMOSPHERIC DEPOSITION

The nine deposition variables were interpolated by cokriging using the multivariate models (with precipitation) at the centres of 10 x 10 km blocks of a regular grid covering all of France. An estimate of the value of atmospheric deposition and the cokriging standard deviation were obtained for each block of the grid.

The size of the neighbourhood for the cokriging was determined by cross validation. The applied neighbourhood contains 32 measurement sites and stations, which corresponds to 48 data values. Thus, the neighbourhood includes the 16 CATAENAT sites closest to the block to be estimated (i.e., 16 values of deposition and 16 values for precipitation) as well as 16 Météo-France stations for which solely values of precipitation are available. No special treatment was applied for Corsica; it was included in the dataset for cokriging, given the observed behaviour for the deposition variograms, it was considered as acceptable.

3.6 MAPPING OF ATMOSPHERIC DEPOSITION

Using the interpolated values (and their standard deviations) mapping is performed for the deposition variables for all of France (Figure 3). The obtained maps show that in most cases, the contribution of the auxiliary variable precipitation brings about an improvement in the deposition maps in terms of: (a) an enhanced level of detail in the description of spatial patterns and (b) a decrease in spatial uncertainty for the cokriged values as is generally attested by the cross validation results (cf. Table 2: lower MSE values for cokriging).

4 Conclusions and perspectives

The applicability of multivariate geostatistical methods was demonstrated for the mapping of atmospheric deposition at the scale of the whole of France. The obtained results are convincing despite the small number (27) of available measurement sites from the CATAENAT network. In particular, the application of cokriging has enabled the production of deposition maps with an improved level of detail in comparison to previous studies.

Spatially structured behaviours were identified and quantified for the nine atmospheric deposition variables analysed at the scale of France. A linear model of coregionalisation was fitted using auxiliary variables in order to characterise the spatial variability of atmospheric deposition. Precipitation was selected, by cross validation, as the auxiliary variable, as it showed a significant contribution for the majority of interpolations by cokriging.

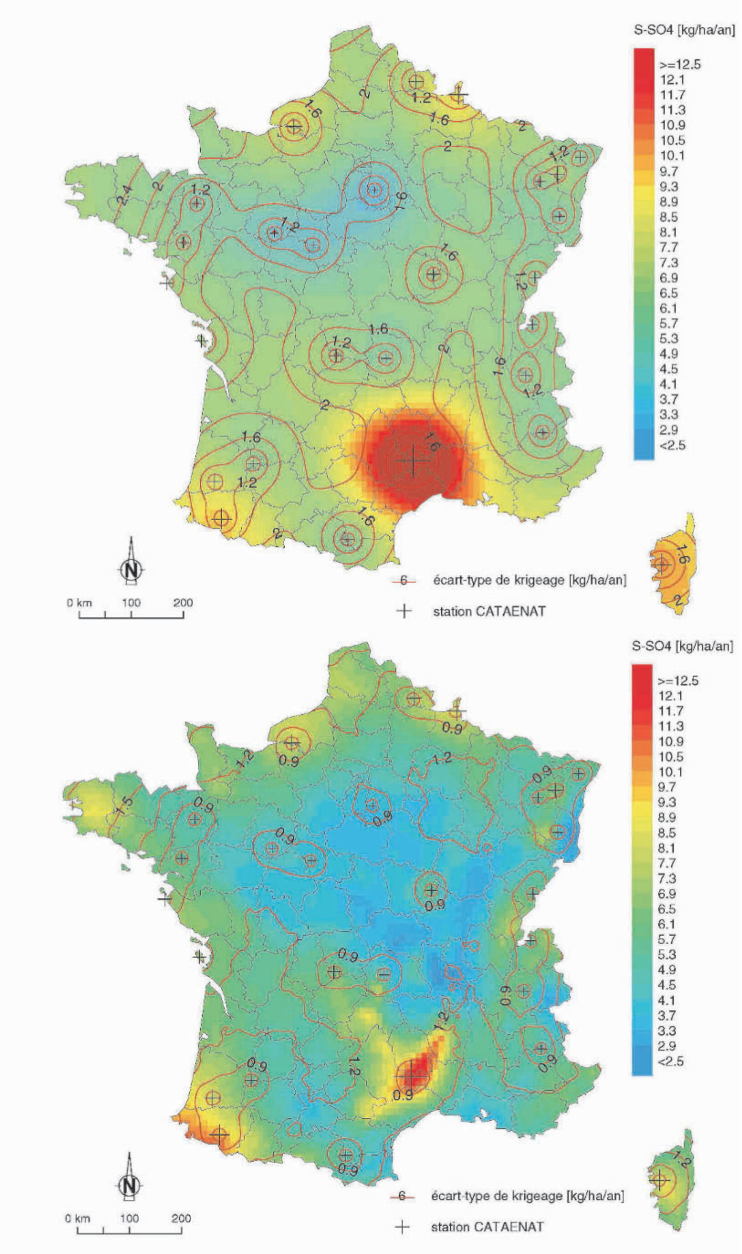


Figure 3. Map of S-SO4 by kriging (above) and map of S-SO4 with cokriging using precipitation (below).

This page intentionally blank

Spatial cross correlations between the different deposition ions were not analysed in this study. However, their contribution should be evaluated in geostatistical terms given the statistical correlation results obtained by Croisé et al. (2002).

Additional meteorological data, either other types of measurements or results from numerical models, could be applied in order to improve mapping of atmospheric deposition for France. Indeed, Geostatistics offer multivariate methods (e.g., the external drift) that are applicable for variables with no common sample locations.

Finally, the issue of exceedance of critical loads for acidity could be answered in probabilistic terms. By using conditional simulation techniques, probability of exceedance, for a given threshold, could be estimated for atmospheric deposition and mapped at the scale of France.

Acknowledgments

The authors would like to thank the 54 foresters who carried out, in a continuous manner, the weekly sampling from 1993 to 1998. Without their constant motivation and effort to contribute to a dataset of high quality, this study would not have been possible.

References

Chilès J.P. & Delfiner P. Geostatistics: modelling spatial uncertainty, Wiley Series in Probability and Mathematical Statistics, 1999, 695 p.

Croisé L., Ulrich E., Duplat P. & Jaquet O. RENECOFOR – Deux approches indépendantes pour l'estimation et la cartographie des dépôts atmosphériques totaux hors couvert forestier sur le territoire français, Editeur : Office National des Forêts, Département Recherche et Développement, 2002, 102 p.

Draaijers G.P.J., Van Leuwen E.P., De Jong P.G.H. & Erisman J.W. Base cation deposition in Europe – Part I. Model description, results and uncertainties, Atmospheric Environment, vol. 31, no. 24, 1997, p. 4139-4157.

Isatis Version 4.0.2, Géovariances, France, 2002.

Johnson D.W. & Lindberg S.E. Atmospheric deposition and forest nutrient cycling, Springer Verlag, 1992, 707 p.

Goulard M. Inference in a coregionalisation model. In: Armstrong M. (Ed.) Geostatistics, Kluwer Academic Publisher, Amsterdam, Holland, 1989, p. 397-408.

Party J.P., Probst A., Thomas A.L. & Dambrine E. Calcul et cartographie des charges critiques azotées en France : application de la méthode empirique, Pollution atmosphérique no. 172, 2001, p. 531-544.

Posch M., De Smet P.A.M., Hettelingh J.P. & Dowing R.J. Modeling and mapping of critical thresholds in Europe, Status report, National Institute for Public Health and the Environment, Netherlands, 2001, 188 p.

Wackernagel H. Multivariate geostatistics, Springer Verlag, 1995, 256 p.

GEOSTATISTICAL AND FOURIER ANALYSIS APPLIED TO CROSS-HOLE TOMOGRAPHY SEISMIC DATA

JORGE M.C.M. CARVALHO and ABÍLIO A.T. CAVALHEIRO
Department of Mining & Geoenvironmental Engineering - CIGAR
Faculty of Engineering - University of Porto
R. Dr. Roberto Frias, 4200-465 Porto - Portugal

Abstract. Geophysical methods, for instance seismic cross-hole tomography, are fundamental investigation tools in geotechnical site characterization. Tomographic inversion procedures allow associating an average P and/or S seismic wave velocity to a set of cells and, by "contraction", to their central points. It is then possible to deduce soil elastic modulii necessary to geotechnical design. Geostatistical tools may be a useful methodology in modeling the spatial variability of such regionalized geophysical and geotechnical variables as well as in their final mapping. However there are various aspects that shall be taken in consideration. In this study geostatistical and Fourier analysis procedures are applied to data resulting from a seismic cross-hole tomography survey using three boreholes defining two sections, in a site of heterogeneous granite weathered profile of Porto area, in the framework of the city metro construction. Mapping of both shear modulus and Poisson ratio, obtained with different interpolation methods and strategies, are presented and discussed, including their related estimation and "contraction" variances. It is also discussed the application of Fourier analysis aiming namely at quantifying increases or decreases of variability due to changes of data frequency content in estimation, as well as an inference of the necessary sampling rate capable of detecting high frequency events, i.e. a measure of the possibility of recovering the "original" regionalized continuous function through bandlimited interpolation, within a certain error, based on Shannon sampling theorem.

1 Introduction

Geophysical site characterization with geotechnical purposes generate spatially distributed discrete data in which are based the final "continuous" maps. Several interpolation methods and strategies may be used to produce such maps/images. It is important to obtain accurate models of the underground reality as well as having the possibility of assessing their degree of accuracy/uncertainty.

Geophysical methods, for instance seismic cross-hole tomography, are fundamental investigation tools in geotechnical site characterization. Soil stiffness properties, namely dynamic shear modulus, G_0, and Poisson ratio, μ may be inferred, under very low strain levels, from measured shear (S) and compressive (P) waves velocities. The G_0 and μ data used in this study were obtained according to the following generic stages: i) application of tomographic inversion procedures to field data acquired in a cross-hole tomography seismic survey, using three boreholes defining two rectangular vertical

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 843–852.

adjacent sections, in a site of heterogeneous granite weathered profile of Porto area, in the framework of the city metro construction (Pessoa, J.M., et al. 2002), ii) partition of the two sections in square or rectangular cells having each one associated an average P or S seismic wave velocity value; iii) contraction of each cell towards its midpoint; iv) soil elastic modulii necessary to geotechnical design are then deduce from P and S wave velocities.

Geostatistical tools may be a useful methodology, namely, in modeling the spatial variability of such regionalized geophysical/geotechnical variables as well as in their final mapping. However there are aspects that shall be taken in consideration like the support and dispersion issue, unusual increase of variability due to support contraction, anisotropic behavior and possible departures from stationarity.

In this study geostatistical and Fourier analysis procedures are applied to such data and different interpolation methods and strategies are herein presented and discussed. Namely it is suggested the application of Fourier analysis aiming at quantifying increases or decreases of variability, i.e. a measure of the possibility of recovering the "original" regionalized continuous function through bandlimited interpolation, within a certain error, based on Shannon sampling theorem (Carvalho, J. and Cavalheiro, A, 2000). Fourier analysis in concert with distribution theory is an adequate mathematical tool to deal with both stationary and non-stationary phenomena.

Two types of G_0 and μ data sets are considered, differing in size: one with all the available values and a partial one with half the values (the odd ones).

Several interpolation algorithms written in Matlab language were used in this study.

2 Exploratory Data Analysis

A previous exploratory data analysis was undertaken in order to characterize the studied parameters based on their respective sample distributions as well as detecting expected existing spatial anisotropies and heterogeneities, trend(s), possible mixed populations and lack of homocedasticity. Data sets with all the values and other partial ones with the odd values will be considered.

Figures 1 and 2 show the location points, for the whole and partial referred data sets, corresponding to the two adjacent tomographic vertical sections defined by three boreholes (one centrally located and the two others on each side of the diagrams).

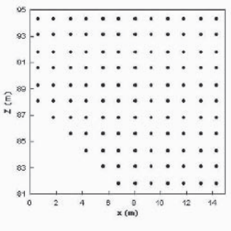


Figure 1. Data location: whole set**.**

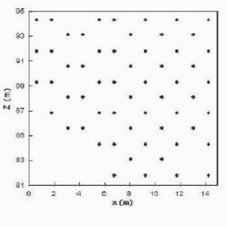


Figure 2. Data location: partial set.

The μ and G_0 distributions are very distinct. The first one (Figure 3) is very negatively skewed and the other (Figure 4) very positively skewed, both showing no apparent mixing of populations.

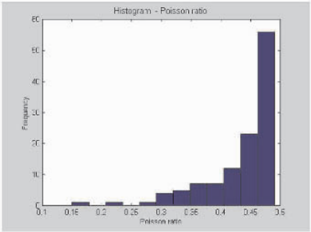


Figure 3. µ: histogram.

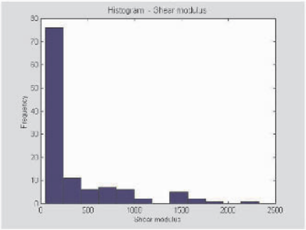


Figure 4. G_0: histogram.

In order to identify and define expected existing trends, four procedures were used: one based on a first order mean increment function d(h) of the vertical lag distance, h, (Chiasson P., 1994). This increment function allows detecting/confirming departures from stationarity: in case of stationarity the data values fluctuate around a mean constant value and the mean increments d(h) will tend to zero; in presence of a trend the increments will tend to increase with lag distance h. For both cases, µ and G_0, there is an increase of the d(h) function, pointing generically to some departure from stationarity (1^{st} order increment function for G_0 in figure5).

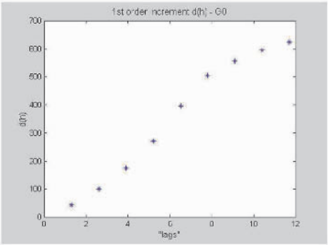


Figure 5. G_0: 1^{st} order increment, d(h).

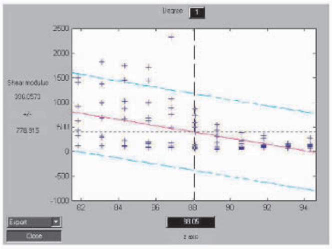


Figure 6. G_0 along vertical axis.

Another procedure consisting in plotting the 2D data values against the coordinates axis (x: horizontal direction; y: vertical direction) and fitting a least square error trend line in the respective scatter plots. In figure 6 there is an example of this last procedure, showing a not very intense trend for G_0, being the one for µ slightly weaker.

A third procedure, using the Matlab function "bootstrap.m", based on the histogram of random correlation coefficient sub-samples between depth z (y axis values) and respective parameter values (example in figure 7, with 500 randomly taken sub-samples). If the minor histogram centre class value is greater than zero, then the relation between the variables should be non fortuitous, which happens to be the case for both studied parameters.

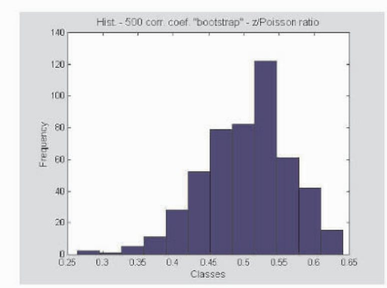


Figure 7. μ: “bootstrap” histogram.

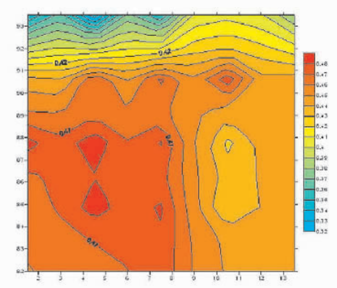


Figure 8. μ: moving window for the mean.

In addition, the moving window statistics procedure, aims at complementing the exploratory analysis namely at detecting heterocedasticity situations. As an example, in figures 8 and 9 can be seen the mean moving window contour plots (using Surfer © software inverse square distance method) respectively for μ and G_0 and in figure 10 the variance moving window contour plot for G_0. This last type of plot can also be used as an additional indicative spatial confidence distribution map for the obtained estimated data.

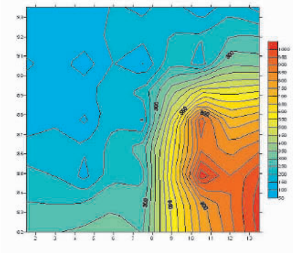


Figure 9. G_0:moving window for the mean.

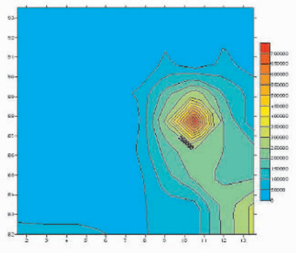


Figure 10. G_0: moving window for the variance.

The results obtained with the partial sets are generically similar, seeming quite representative samples of the whole sets.

3 Structural Analysis

The structural analysis study allowed characterizing the pattern of directional variability of both studied parameters, showing again their distinct behavior. The variogram modeling was done using the software Variowin (Pannatier Y., 1996) along directions 0º (horizontal), 45º, 90º (vertical) and 135º with a 22.5º and 90º (omnidirectional variogram) tolerance angles.

In the case of μ there is no evidence of presence of a trend. The ellipse of geometric anisotropy is depicted by the respective range rose diagram (figure 11) and the omnidirectional variogram is in figure 12.

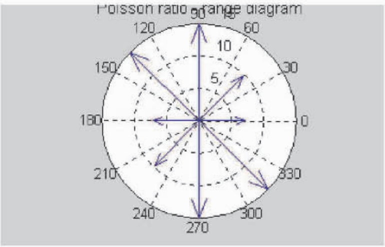


Figure 11. μ ranges rose diagramme.

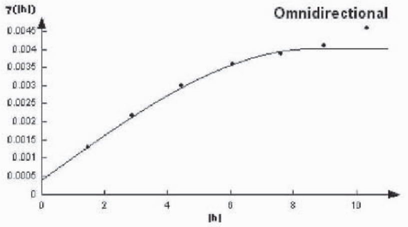


Figure 12. μ omnidirectional spherical variogram model**.**

The G_0 pattern of variability with direction is much less smooth as it can be observed in figures 13 to 16. In fact there are two unbounded variograms along directions 90º and 135º (power model) and two other bounded ones along directions 0º and 45º (spherical model).

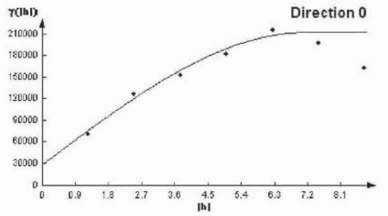


Figure 13. G_0: 0º spherical model.

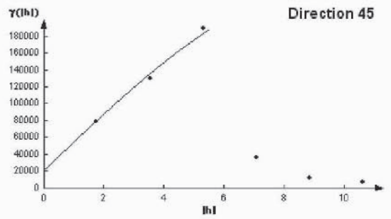


Figure 14. G_0: 45º spherical model.

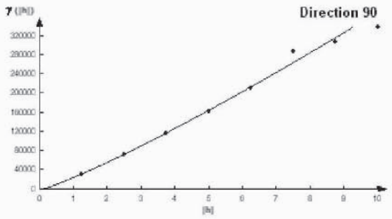


Figure 15. G_0: 90º spherical model.

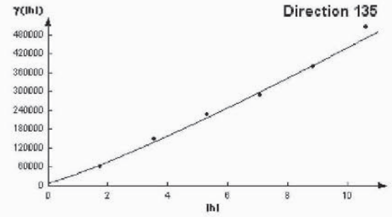


Figure 16. G_0: 135º spherical model.

Along direction 45º there is a discontinuity in the experimental variogram, around 5.5m lag, separating a higher and increasing initial part from a lower decreasing zone. This fact can be understood by looking at figures 25 and 26 showing G_0 estimation maps and at the h-scatterplots of G_0 values separated by lag 7m, along directions respectively 45º and 135º, in figures 17 and 18.

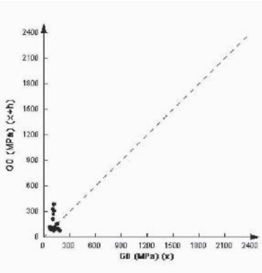


Figure 17. G_0: h-scatterplot, direction 45º, lag 7 m.

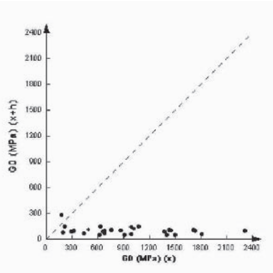


Figure 18. G_0: h-scatterplot, direction 135º, lag 7 m.

The respective variogram surface shows a clear anisotropic behavior (figure 19). However, the G_0 omnidirectional variogram (figure 20) happens to be bounded and very smooth.

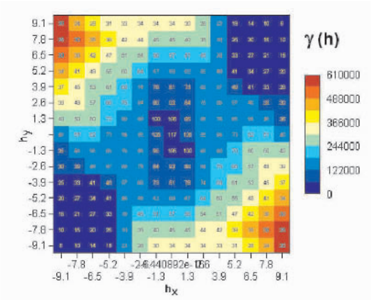


Figure 19. G_0: variogram surface.

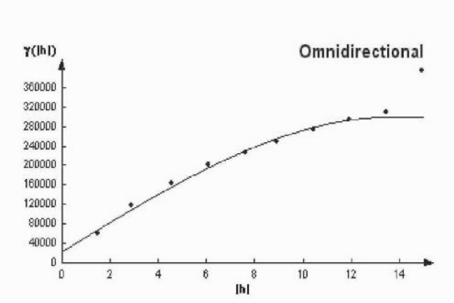
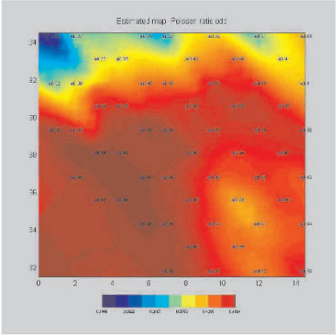


Figure 20. G_0**:**.omnidirectional spherical variogram model.

Based on the obtained generically well behaved variograms and the posterior cross validation and estimation results, it was decided not to try to normalize the data as well as not removing the identified weak G_0 trend.

4 Cross Validation

In order to evaluate the adequacy of the spatial variability models as well as the quality and characteristics of the estimation procedures, three types of cross validation were used, one more conventional and the others based on the referred data splitting into two data subsets. One of the subsets was used to estimation and the other for error control, after estimating globally their values given the respective locations as if they were unknown or to estimate a grid identical to the one obtained with the whole set. When there is enough data this is obviously a practical way to assess the relative quality of the estimation procedures, namely the associated error distributions. In figure 21 can be seen an example of this last cross validation procedure, estimating by ordinary kriging with a geometric anisotropy variogram model (OK) the whole grid with the odd values subset and in figure 22 the histogram of the difference between this last grid values and those obtained using the whole data set. The mean error is 0.0003 and the median 0.0017. The analyzed frequency content of the two grids is very similar.



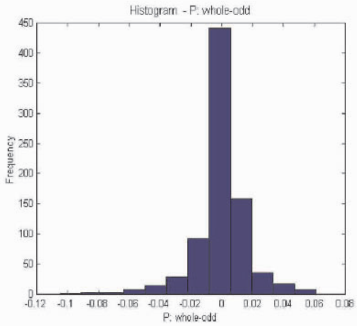


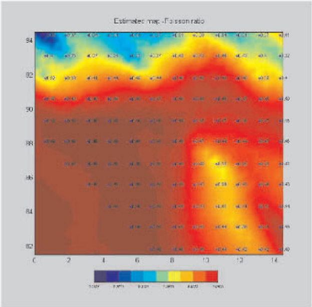
Figure 21. μ: OK estimation map, using odd subset.

Figure 22. μ: histogram of the grid differences.

5 Estimation

The grid estimation, for both parameters, has been achieved using the inverse square distance method (herein not discussed), isotropic and anisotropic ordinary kriging (OK) and kriging with a trend (KT) and some explorative experiences with sinc bandlimited interpolation based in Shannon sampling theorem (Papoulis, A., 1962).

The present type of situation, regarding the way of obtaining data, implies a change of support from the original square or rectangular cells to their midpoints. This fact implies a combination of an increase and a decrease of variance resulting from the kriging estimation procedures due, respectively, to the decrease of support ("contraction" variance) and to the well-known smoothing (high cut filtering) effect of kriging.



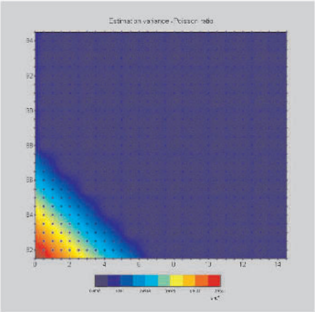


Figure 23. μ: OK estimation map, using the whole data set.

Figure 24. μ: OK variance map.

In figure 23 is the μ OK map of estimates using the whole data set (shown in the figure) and in figure 24 the respective kriging variance map. As expected, the higher kriging variance values are positioned in the lower left corner of the map in the non sampled zone. Otherwise, as a measure of uncertainty, is not very valuable information.

Figures 25 and 26 show the G_0 estimated maps using respectively kriging with a vertical trend model (estimates: mean = 336, range of values =1968; standard deviation = 406.5) and ordinary kriging (estimates: mean = 336, range of values = 1400; standard deviation = 344). The first option is advantageous considering the larger range of values (lesser smoothing effect) and appeared more realistic when confronted with the available local borehole geological information.

Fourier analysis may be an interesting alternative way to understanding and characterizing the observable changes in the estimation maps resulting from different interpolation procedures, as it will be discussed in item 6.

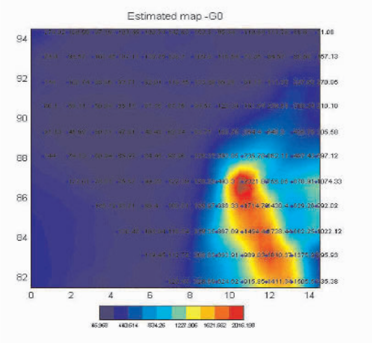


Figure 25. G_0: KT estimation map, using the whole data set.

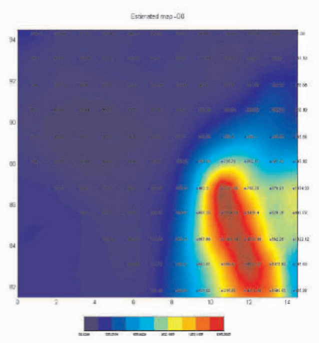


Figure 26. G_0: OK estimation map, using the whole data set.

6 Fourier Analysis

The possibility of accurately reconstructing a sampled spatial and/or time "continuous" signal depends greatly on the adequacy of the sampling rate to the frequency content of the studied system or in other words to its variability rate. In addition it is also important to use interpolating procedures convenient in terms of the estimated data frequency content. Bandlimited interpolators may be an interesting alternative to other methods providing the data is abundant enough and supposedly regularly sampled.

In figures 27 and 28 can be seen the spectrograms (instant spectra) of the estimated G_0 grid values using respectively KT (figure 27) and OK (figure 28). Each vertical coloured column cells corresponds to the related space (horizontal axis) window frequency content (instant spectra). The analyzed estimated values vectors correspond to sequential grid lines having the same length as the Hanning space window filter used to build the spectrograms. This way it is possible to identify, located in space, changes in the frequency content among different estimated grids. In figure 29 are plotted the more conventional Fourier spectra corresponding to the same referred two data vectors (very similar frequency content). This last frequency function, giving only the overall data frequency content, does not permit to combine spatial and frequency information as the spectrogram does but, even so, may be an useful informative tool for data and sampling rate characterization.

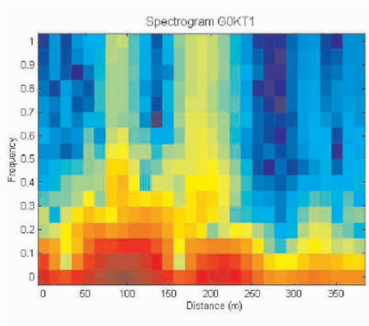


Figure 27..G_0: spectrogram, KT.

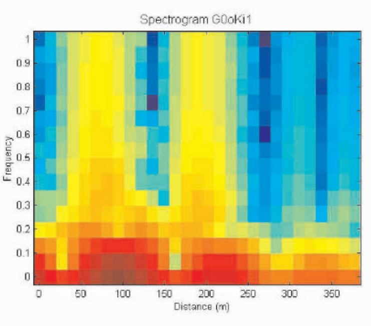


Figure 28. G_0: spectrogram, OK.

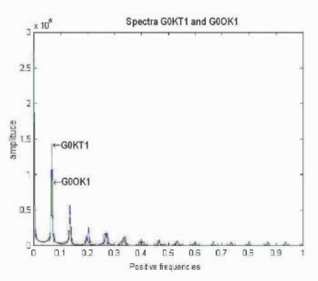


Figure 29. G_0: KT and OK estimates amplitude spectra.

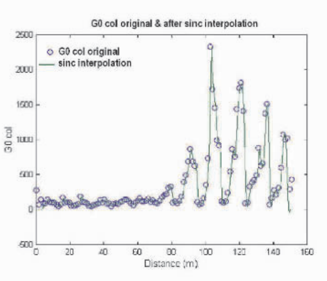


Figure 30. G_0: partial odd set and sinc bandlimited interpolated result.

In both figures 30 and 31 can be seen two graphs corresponding respectively to the partial odd and whole G_0 data sets and the obtained sinc (bandlimited) interpolated data along matrix columns having for each case twice the number of values. In figure 32 are the Fourier spectra plots of the last two referred data sets. Excepting the edges the differences between the graphs are almost indiscernible. Among others, these results are quite promising in both space and frequency domain characteristics.

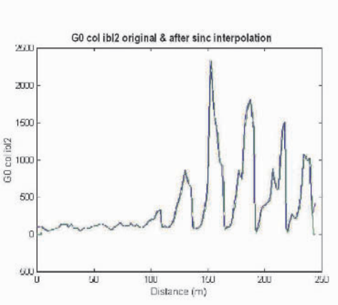


Figure 31. G_0: whole set and sinc bandlimited interpolated almost indiscernible graphs.

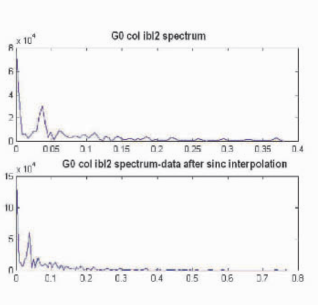


Figure 32. G_0: Fourier spectra of whole set (above) and of sinc interpolated set (below).

7 Final Notes

The present study basically explores combines and compares different tools and approaches used to obtain final images in the framework of geophysical/geotechnical site characterization.

One of the main generic conclusions appears to be the advantageous combination of (geo)statistical tools with Fourier analysis in order to gain a better understanding of the processes underlying estimation namely in terms of assessing their related characteristics and accuracy/uncertainty.

The adequacy of the sampling rate to the variability (frequency content) of the studied system is a key point closely related to the possibility level of accurately recovering a continuous entity from a discrete sample. Fourier analysis may be, also in this context, a very informative tool.

In particular, may be mentioned the following facts: i) the structural analysis worked very well with the seismic derived data; ii) the obtained graphical output maps based in previously kriged grids followed by lowpass interpolation using the Matlab "interp.m " function appears to be of very good quality; iii) the sinc bandlimited interpolation seems to be a promising alternative method namely in terms of frequency content preservation. It is also worth mentioning the instant spectrum, relating space and frequency content information, as an enlightening tool namely in characterizing interpolated grids as well as in the understanding of the underlying studied geo-systems.

Acknowledgements

We would like to acknowledge "Normetro, Agrupamento do Metropolitano do Porto ACE", and "Transmetro – Construção de Metropolitano, ACE", for allowing using the present data.

References

Bracewell, R.N., *The Fourier Transform and its Applications*, McGraw-Hill, 1978.

Carvalho, J.M.C, *Modelação e Tratamento Geoestatístico de Dados SPT e Sísmicos.* PhD thesis, FEUP, 2002.

Carvalho, J.M.C. and Cavalheiro, A., *Geostatistics Applied to SPT Data – A Case Study.* Geostatistics 2000, Cape Town, Vol 2", W.J. Kleingeld and D.G. Krige, Editors, 2001.

Chiasson, P. Lafleur, J, Soulié, M., Law,T., *Characterizing Spatial Variability of a Clay by Geostatistics.* Can. Geotch., Vol. 32, Nº 1, 1995.

Deutsch, C.V. and Journel, A.G., *GSLIB: Geostatistical Software Library and User's Guide*, 2nd Edition, Oxford University Press, 1998.

Goovaerts, P., *Geostatistics for Natural Resources Evaluation*, Oxford University Press, 1997.

International Society for Rock Mechanics (1988). *Suggested methods for seismic testing within and between boreholes.* Int. J. Rock Mech. Min. Sci. and Geomech. Abstr., 25 (6), pp. 447–472.

Isaaks, E.H. and Srivastava, M.R., *An Introduction to Applied Geostatistics*, Oxford University Press, 1989.

Matheron, G, *Traité de Geostatistique Appliquée.* tome I, Mémoires du B.R.G.M., Nº14, 1962.

Pannatier, Y., *Variowin – Software for Spatial data Analysis*, Springer, 1996.

Papoulis, A., 1962, "*The Fourier Integral and its Applications*", McGraw-Hill", Publ. Comp., 1962.

Pessoa, J.M., *Aplicação de técnicas tomográficas à prospecção sísmica entre furos de sondagem.* Trabalho de síntese, Provas de Aptidão Pedagógica e Capacidade Científica. Universidade de Aveiro, 1990.

Pessoa, J.M., Carvalho, J.M.C.; Carminé, P., Fonseca, A.V., *Tomografia de velocidade de ondas sísmicas S e P numa zona de alteração de granito do Porto*. 8º Cong. Nac. de Geotecnia. SPG, Lisboa, 2002.

Viana da Fonseca, A. et al. *Ensaios sísmicos entre furos em perfis de alteração de granito do Porto – determinação de parâmetros elásticos e correlações*. 8º Cong. Nac. de Geotecnia, 2002.

THE IMPORTANCE OF DE-CLUSTERING AND UNCERTAINTY IN CLIMATE DATA: A CASE STUDY OF WEST AFRICAN SAHEL RAINFALL

ADRIAN CHAPPELL and MARIE EKSTRÖM
School of Environment & Life Sciences, University of Salford, Manchester, UK.
Climate Research Unit, University of East Anglia, Norwich, UK.

Abstract. Climate station data are increasingly being required in gridded formats for many purposes in applied and theoretical environmental science. However, spatial analyses of climate data are particularly dependent on the location of the samples. Rainfall station data at 123 locations for 1943 in the west African Sahel were used as a case study to investigate the influence of station location on estimates. The rainfall data were de-clustered to remove the influence of their preferential location in the wet, western coastal area. The mean annual rainfall for this year of data and for that of years between 1931 and 1990 were considerably smaller than the simple arithmetic mean of the data. An omni-directional variogram and indicator variograms for several thresholds were computed and fitted with conventional models. The rainfall data, de-clustered histogram and variograms were used to condition simulated annealing realizations. The simulations were compared with maps of rainfall produced with raw data using simulated annealing and ordinary kriging, separately. The results showed that in addition to the improved representation of variability in rainfall, the use of simulated annealing with de-clustered rainfall data provided a new insight into the magnitude and spatial distribution of rainfall in the region.

1 Introduction

Global analyses of climate change impacts (e.g. Hulme *et al*., 1999) and temporal variation in global and regional climates (Jones and Hulme, 1996; Dai *et al*., 1997) make use of a considerable body of climate station sample data. These data are also increasingly being required in spatially complete (gridded) formats for many purposes in applied and theoretical environmental science (New *et al*., 2002). However, spatial analyses of climate data are particularly dependent on the location of the samples. For example, samples of rainfall (stations) over the west African Sahel (WAS) are typically very sparse in some places and dense in others because the network as a whole is dependent on each country's decisions on the fate of their stations (Figure 1). Furthermore, estimates made at unsampled locations are complicated by the location of the existing samples in the established climatology of a wet, western coastal and dry, eastern continental and northern sub-Saharan areas. Perhaps as a consequence of the relatively few data or the need for relatively rapid interpolation techniques (e.g., splines) to process climate data across the globe, the use of geostatistics has not been widely applied in global climatological datasets. Thus, until recently the temporal pattern of rainfall was established by a simple arithmetic average of the annual rainfall totals. In

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 853–863.

the WAS these totals were dominated by rainfall in June, July and August (Figure 2). More recent work has attempted to account for the location of rainfall station locations in high- and low-valued regions by subtracting the rainfall value from the long-term station mean to produce anomalies for which the average is obtained using a distance weighting scheme (Jones and Hulme, 1996; Dai *et al.*, 1997). The simple arithmetic average is likely to be highly susceptible to the clustering of samples, particularly in the high-valued western region of the WAS. Because of its variable (location-dependent) filtering property, the appearance (smoothness) of kriging maps depends on the local data configuration. For irregularly spaced data, the map is more variable where sampling is dense than where it is sparse (Isaaks and Srivastava, 1989). Such an effect may create structures that are pure artefacts of the data configuration. One solution consists of utilizing simulation algorithms which as opposed to kriging algorithms reproduce the full covariance everywhere and represent the characteristics of the data and its inherently local variability.

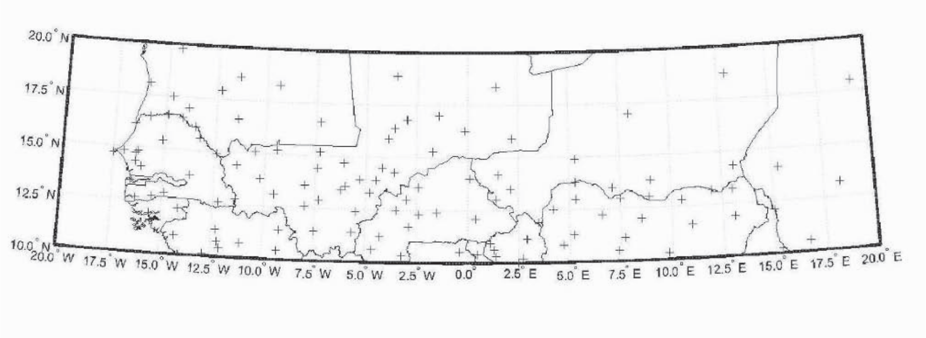


Figure 1 Rainfall station locations in the west African Sahel in 1943.

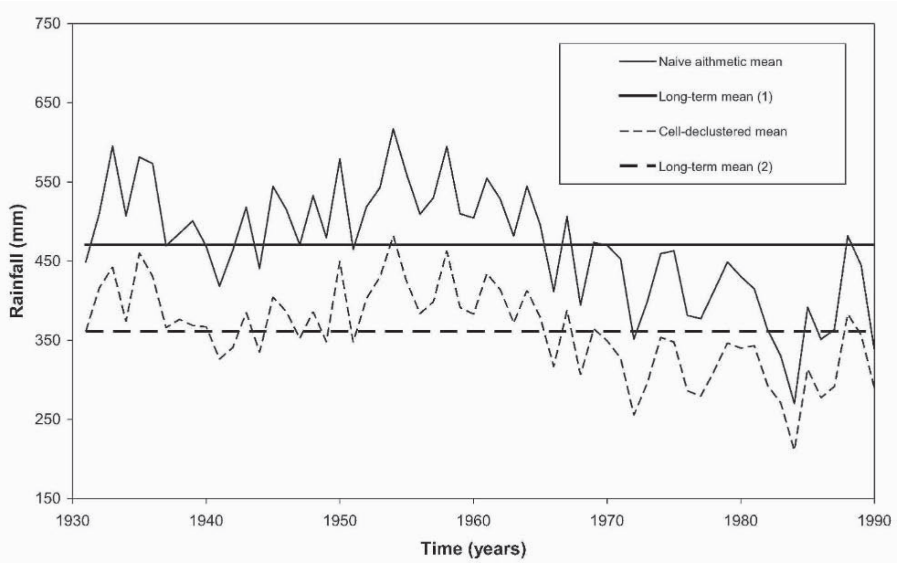


Figure 2 Regional statistics of annual rainfall for the west African Sahel.

The aim here is to demonstrate the influence of the rainfall station locations and their clustered nature on a map of rainfall and to offer a solution that takes into account de-clustered data. A single year of rainfall data (123 locations) for the WAS will be used as a case study. The first objective is to apply cell de-clustering to the univariate rainfall distribution and to characterise the change in the simple arithmetic mean. Secondly, a rainfall map will be produced by kriging with an omni-directional variogram and compared with a realization from simulated annealing conditioned on the clustered histogram, the previous primary variogram and indicator variograms. Finally, realizations will be produced using the same simulation approach but with thresholds for the indicator variograms transposed from the de-clustered histogram.

2 Methods

This section provides a brief outline of each of the techniques used here. All are common in the field of geostatistics and further details can be found elsewhere. The originality of the paper lies in the combination of methods and its timely application to climate data and an innovative transform between the sample cumulative distribution function (CDF) and the de-clustered CDF that enables simulated annealing to use the thresholds of the indicator variograms for the de-clustered distribution.

2.1 CELL DE-CLUSTERING

The study area was divided into rectangular or square cells and the number of cells with at least one datum was counted, and so too were the number of data in each cell. Each datum location received a weight $\lambda=1\ /\ (B\ .\ n_b)$, where B is the number of cells that contains at least one datum and n_b is the number of data within each cell (Goovaerts, 1997, p. 81). The weight gave more importance to isolated locations (Isaaks and Srivastava, 1989). Several cell sizes and origins were tried to identify the smallest declustered mean because large values of rainfall were preferentially sampled. Erratic results caused by extreme values falling into specific cells were avoided by averaging results for several different grid origins for each cell size. The de-clustering procedure implemented in GSLIB was used (Deutsch and Journel, 1998).

2.2 OMNI-DIRECTIONAL VARIOGRAM AND INDICATOR VARIOGRAMS

Latitude and longitude data are not linear measures of distance, especially over large areas. However, close to the equator the distortion is very small and assumed to have a negligible effect on the variography and mapping. Despite strong gradients in rainfall across longitudes and latitudes there were insufficient data to reliably quantify this anisotropic spatial variation (Webster and Oliver, 1992). Since, the inter-annual variation in rainfall is large there is no basis to use data from other years to better inform the variogram and modelling. Consequently, the traditional variogram of rainfall was calculated in all directions (omni-directional). There was a strong possibility that the systematic change in rainfall across the WAS would induce a trend in the variogram. To investigate this possibility and to characterise the spatial structure for different magnitudes of rainfall, indicator variograms were calculated. Five percentiles of the raw rainfall cumulative distribution function (CDF) (Table 1) were selected and used to convert the rainfall at all locations into five indicator sets of data. Values of rainfall that exceeded a threshold were assigned a value of 0 whilst all others were set to 1. Those thresholds for the raw rainfall CDF were transposed to the de-clustered rainfall CDF to provide five equivalent thresholds (Table 1) for use in the simulated annealing. Several conventional models (spherical, exponential, Gaussian and power) were fitted to the experimental variograms using weighted least squares in Genstat (Genstat 5 Committee, 1992). Selection of the best fitted model was based on the smallest square-root of the average difference between the observed and predicted values (RMSE).

Table 1 Model parameters fitted to the omni-directional variogram and indicator variograms for selected rainfall thresholds.

Percentile (%)	Omni	10	25	50	75	90
.aw threshold (mm)	-	178.8	302.8	470.0	660.5	930.0
Model fit (RMSE*)	4171.0	0.01	0.01	0.01	0.01	0.01
Model	Gauss	Gauss	Sph	Exp	Exp	Gauss
Range (a)	$^{\#}3.5$	$^{\#}4.60$	6.89	$^{+}2.37$	$^{+}1.82$	$^{\#}2.71$
Spatially dependent sill (c)	97537.0	0.09	0.19	0.27	0.21	0.07
Nugget (c_0)	11186.0	0.01	0.01	0.02	0.00	0.02
)e-clustered threshold (mm)	-	218.3	472.1	709.2	872.1	1036

*RMSE – Square root of the average of the squared difference between the observed and predicted values $^{+}$Effective range is 3a. $^{\#}$Effective range is 95% of its sill variance.

2.3 KRIGING AND SIMULATED ANNEALING

Using GSLIB (Deutsch and Journel, 1998), the parameters of the models fitted to each of the omni-directional and indicator variograms of rainfall were used to solve isotropic ordinary punctual kriging equations and estimate rainfall every 0.5 degree across the WAS. Isarithmic lines were threaded through these estimates.

Simulated annealing (SA) was performed using GSLIB on several combinations of simulation conditions to investigate the effect of each on the resulting realization. In all cases the raw data and the histogram were used to condition the simulation. Therefore, the SA was conducted separately on the raw rainfall histogram and that of the de-clustered histogram. In the first case, the omni-directional variogram was included in the conditional simulation. In the second case that variogram was replaced with the indicator variograms. Therefore, the SA was conducted separately on the thresholds for either the raw rainfall histogram or those of the de-clustered rainfall histogram. In the final case, the indicator variograms and the omni-directional variogram were combined. The sill of the omni-directional variogram was standardized to the variance of the univariate distribution to ensure that the variance of the initial random image matched the spatial (dispersion) variance implied by the variogram model (Deutsch and Journel, 1998; p. 187).

3 Results

The simple arithmetic mean rainfall for the WAS during 1943 was 518 mm. The minimum de-clustered mean rainfall was 385 mm for cells of 12 degrees across longitudes and of 6 degrees across latitudes (Figure 3). To demonstrate the importance of this procedure and to place this difference in mean annual rainfall into the context of long-term data for the same region, the de-clustered mean was calculated for annual data between 1931 and 1990. The results are included with the other aggregated statistics in Figure 2. The CDFs for the raw data and the de-clustered data are shown in Figure 4. The omni-directional and indicator experimental variograms for raw and de-clustered rainfall thresholds are shown in Figure 5. Models that fitted the data best in the least-squares sense are also shown in Figure 5 and the model parameters are

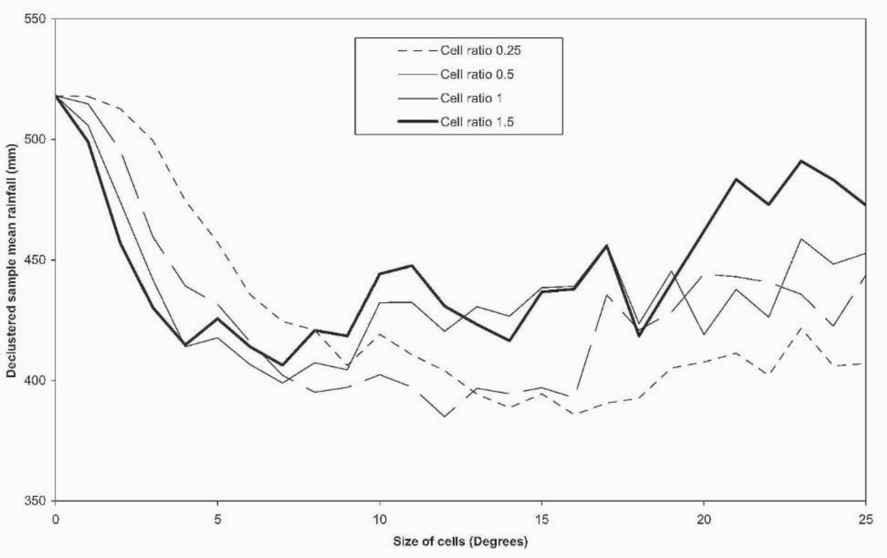


Figure 3 Results of the de-clustering procedure using different ratios for cells.

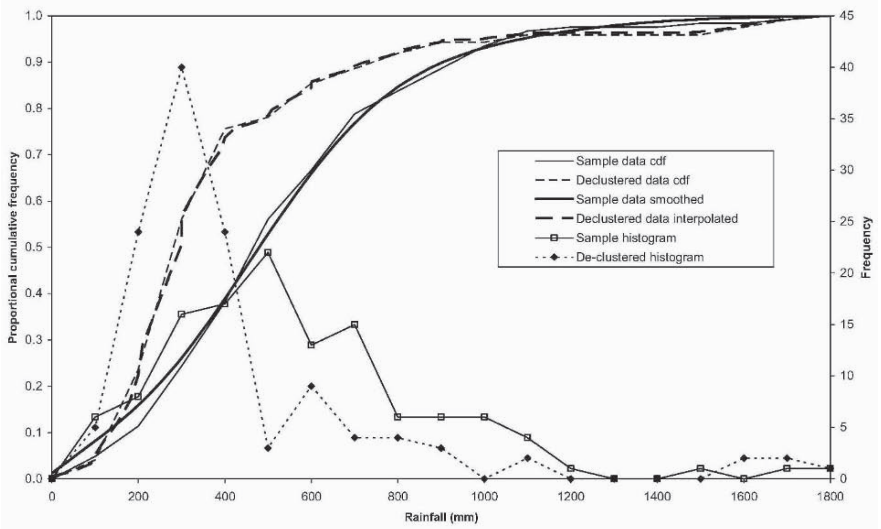


Figure 4 Cumulative distribution functions and histograms for raw and de-clustered rainfall data.

in Table 1. The Gaussian model was the only one fitted to the omni-directional variogram that had a positive nugget value. In general, the models had unexplained variation (nugget variance) that was only a small proportion of the sill variance. The model fitted to the median rainfall variogram had the largest range and spatially dependent variance and those fitted to variograms of smaller rainfall thresholds had smaller values in the respective parameters. This pattern was replicated in the models fitted to variograms of large rainfall thresholds.

Figure 5 Experimental variograms and fitted models for raw rainfall data

The spatial location of the rainfall stations and the total annual rainfall at each station is represented by the size of the symbols in Figure 6a.

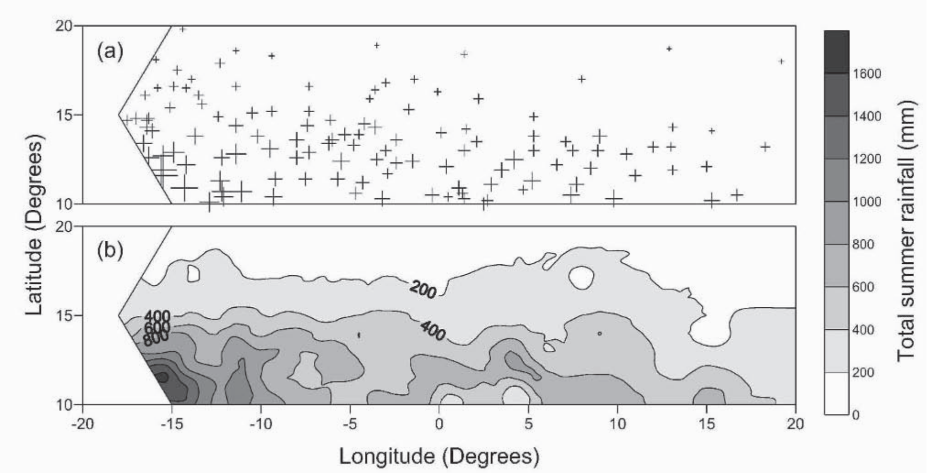


Figure 6 Location of the rainfall stations (a) used with ordinary kriging to estimate total annual summer rainfall every 0.5 (b) in the WAS.

The ordinary kriging map of rainfall displays a smooth and continuous surface of rainfall across the region. It represents the high magnitude of rainfall in the south-west corner and the low magnitude of rainfall in the east and northern parts of the region. It is notable that areas where there are more rainfall stations are more variable than those with few stations. The maps of simulated annealing for raw data and de-clustered data are shown in Figures 7 and 8, respectively.

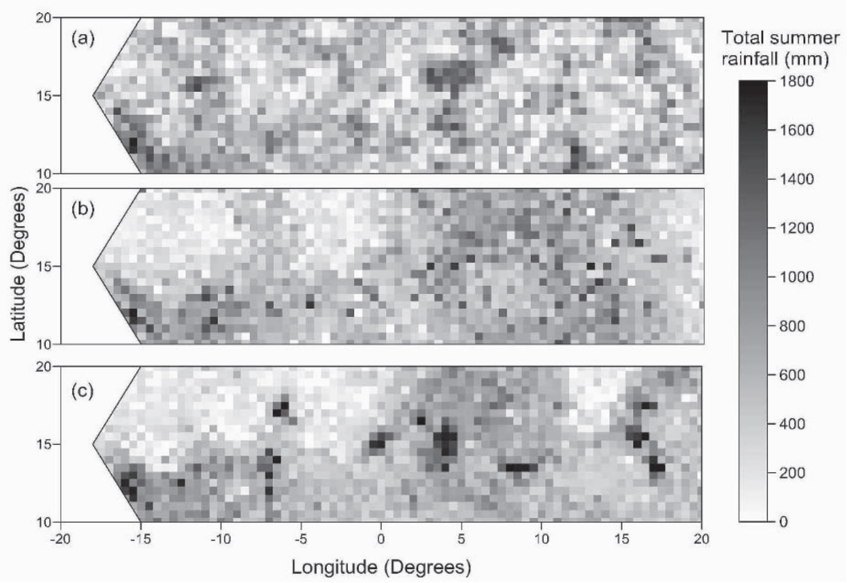


Figure 7 Maps produced by simulated annealing every 0.5° conditioned with raw data, its histogram and separately by (a) omni-directional variogram, (b) indicator variograms for thresholds of the histogram and combined in (c).

Figure 7 shows a magnitude of rainfall that is considerably larger than that evident in Figure 8. The former figure displays spatial structure that resembles more closely the map of rainfall produced by ordinary kriging than the latter. Notably, the maps in Figure 8 (b and c) show a consolidation of the largest rainfall in the south-west which is in marked contrast to the much larger spatial distribution of large rainfall in Figure 7. In both figures the first panel (a) represents the simulation conditioned with the omni-directional variogram. The second panel in both figures (b) shows the results of the being conditioned with the indicator variograms. The final panels (c) in the figures show the simulation conditioned by the omni-directional and indicator variograms. In both cases (Figure 7a and 8a) the omni-directional variogram has provided a general level of structure to the data. However, there remains a considerable amount of variability in the rainfall pattern. In contrast, the indicator variograms in Figures 7b and 8b provided a clear structure with much less variability. The combination of omni-directional and indicator variograms (Figure 7c and 8c) reduced the rainfall variability further and this is particularly evident at the extremes of the rainfall spectrum.

4 Discussion

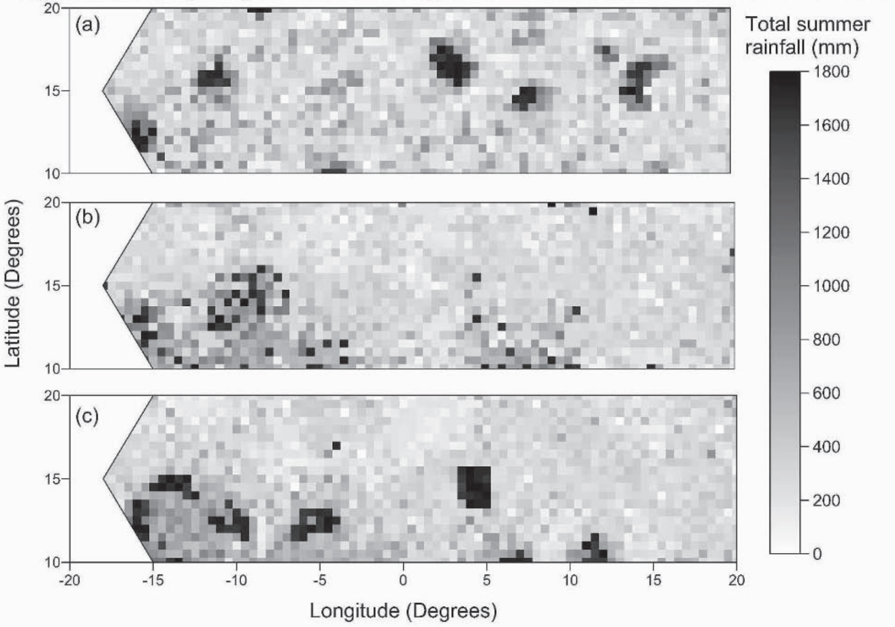


Figure 8 Maps produced by simulated annealing every 0.5° conditioned with raw data, the de-clustered data histogram and separately by (a) the omni-directional variogram, (b) indicator variograms for thresholds of the de-clustered histogram and combined (c).

The large difference (25%) between the simple arithmetic mean of the raw rainfall data for 1943 and that of the de-clustered data characterises the clustered nature of the rainfall station locations. This finding is hardly surprising since the stations were not designed collectively to represent the region. The impact of de-clustering for this one year of data is placed into context when all years of rainfall data (1931-1990) for the region are de-clustered and compared with the simple arithmetic mean and a recent area-weighted, inverse-distance average statistic (Figure 2). It appears that the downturn in rainfall since 1970 is much less reduced in the de-clustered time series in comparison with the other statistics. This crude analysis does not account for the changing location of the rainfall station network over time which has been recently shown to explain the majority of the downturn (Chappell and Agnew, 2004). The histogram of the raw rainfall data is approximately normally distributed (skewness 1.17) whilst that of the de-clustered rainfall data is much more positively skewed (2.54). The univariate distribution of rainfall is notoriously positively skewed as a consequence of relatively few high magnitude rainfalls in time and space. This contrasts markedly with the much larger number of rainfall amounts that are between small to medium in their magnitude. Thus, it appears that the clustered nature of the rainfall station locations, close together in the wet region, has hidden the true and expected strongly positively skewed distribution. This distribution only became evident after the de-clustering procedure.

The location of the rainfall stations also influence the interpolated rainfall map produced by ordinary kriging (OK). This is evident in the north-east part of the map where few stations are located and the spatial variation of the rainfall is very small. Given the ‘spottiness’ of semi-arid rainfall (Sharon, 1972) the variation in this area is likely to be similar, if not greater, than that found in the wetter, coastal region where many more stations are located. The location and magnitude of the rainfall stations cause the interpolated map to have the well-established latitudinal bands of rainfall across the WAS (Figure 6). It is very likely that alternative interpolators will reproduce a very similar pattern of rainfall as that produced by ordinary kriging.

The maps of rainfall produced using SA conditioned with the raw sample data (Figure 7) show considerably more variation throughout the region than that evident in the OK map. These maps display some of the spatial structure of rainfall that was evident in the OK map. Generally, there is more rainfall in the lower latitudes than in the higher latitudes. However, there remains a large amount of rainfall (ca. 1200-1400 mm) in the north-east quadrant of the map. This feature is not consistent with the area being reputedly the driest in the region. Furthermore, the rainfall amounts simulated in this quadrant are similar to those evident across the longitudes between 10-15° latitude. This pattern is also unrealistic because it contradicts the accepted climatology for the region where the largest rainfall is found in the south-west quadrant. In contrast, the maps of rainfall produced using SA conditioned with the de-clustered rainfall data exhibit these desirable climatological characteristics (Figure 8). The north-eastern quadrant has rainfall that varies between 0 mm and 400 mm but which is predominantly at the lower end of that range. The rainfall in the south-western region varies between 200 mm and 1800 mm but is dominated by amounts greater than ca. 800 mm.

Since the omni-directional and indicator variogram are the same in both sets of maps the single most important factor controlling the spatial distribution of the SA is the histogram. It is evident from the distribution of the rainfall in the maps of Figure 8 that the de-clustered histogram conditions the simulation to provide better realizations than those using raw data. Of secondary importance to the realistic nature of the simulation appears to be the use of indicator variograms. The indicator variograms appear to structure the spatial distribution of rainfall at each of the thresholds. On its own the omni-directional variogram does not have the same impact on the simulation. However, it appears to make an important contribution when combined with the indicator variograms. Its impact is to cluster in space similar rainfall amounts and reduce the variability in the map (Figure 7c and 8c). This is probably a result of the Gaussian model which has a much larger influence at small lags than at larger lags.

5 Conclusion

The map of rainfall for the west African Sahel (WAS) produced using ordinary kriging (OK) displayed a smooth and continuous surface of rainfall across the region. It represented the conventional understanding of the spatial distribution of rainfall in the region. Simulated annealing (SA) offered an alternative representation of rainfall in the region by emphasising the characteristics of the data, the histogram and the spatial structure. There were important similarities between the SA map and the OK map of rainfall. However, the realizations of the former were largely unacceptable because the fundamental characteristics of the regional climatology were not displayed. It is likely that the realizations could be improved with the inclusion of, for example, soft ancillary information about two latitudinal classes.

Of probably greater significance to the use of gridded estimates than the difference between OK and SA maps is the importance of de-clustering the rainfall data. De-clustering of the rainfall data made a considerable difference to the univariate histogram and to the mean annual rainfall for 1943. When that rudimentary analysis was applied to annual rainfall between 1931 and 1990 the time series showed a much reduced downturn in WAS rainfall which opposed the conventional understanding, but supported recent developments in the importance of rainfall station locations on the aggregated regional statistics (Chappell and Agnew, 2004). The de-clustered weights cannot easily be used in the OK map without interfering with the unbiasedness conditions for kriging. However, the inclusion of the de-clustered histogram in the SA map of rainfall produced realizations (using either the indicator variograms alone or combined with the omni-directional variogram) that were consistent with the regional climatology, but radically different to that of the OK map of rainfall and therefore of the accepted spatial distribution of rainfall for the region.

The SA map of de-clustered rainfall across the WAS provided a new insight into the distribution of rainfall. This analysis suggests that despite the well-known declustering effect of kriging it could not capture the realistic distribution of rainfall because of the inherently extreme nature of the clustered climate station data. Furthermore, the uncertainty in the data is not considered. Thus, SA offers considerable advantages over this traditional approach.

Acknowledgements

The authors are grateful to Mike Hulme formerly of the Climatic Research Unit, University of East Anglia for providing the data used in this analysis and to the anonymous reviewers whose comments did much to improve the manuscript.

This page intentionally blank

References

Chappell, A. and Agnew, C.T., Modelling climate change in west African Sahel rainfall (1931-1990) as an artifact of changing station locations. *International Journal of Climatology*, vol. 24**,** 2004, p. 547-554.

Dai, A., Fung, I. Y. and Del Genio, A. D., Surface observed global land precipitation variations during 1900-88. *Journal of Climate*, vol. 10**,** 1997, **p.** 2943-2962.

Deutsch, C.V.and Journel A.G., *GSLIB Geostatistical Software Library and User's Guide*. Oxford University Press, Oxford, 1998.

Genstat 5 Committee, *Genstat 5, Release 3, Reference Manual,* Oxford University Press, Oxford, 1992.

Goovaerts, P. , *Geostatistics for natural resources evaluation*. Oxford University Press, Oxford, 1997.

Hulme, M., Mitchell, J.F.B., Jenkins, J., Gregory, J.M., New, M., Viner, D., Global climate scenarios for fast-track impacts studies. *Global Environmental Change Supplement*, Iss:S3-S19, 1999.

Isaaks, E.H. and Srivastava, R.M., *Applied Geostatistics*. Oxford University Press, Oxford, 1989.

Jones, P. D. and Hulme, M., Calculating regional climatic time series for temperature and precipitation: methods and illustrations. *International Journal of Climatology*, vol. 16, 1996, p. 361-377.

New, M. Lister, D., Hulme, M. and Makin, I., A high-resolution data set of surface climate over global land areas. *Climate Research*, vol. 21, 2002, p. 1-25.

Sharon, D., The spottiness of rainfall in a desert area. *Journal of Hydrology*, vol. 17, 1972., p. 161-175.

Webster, R, and Oliver, M.A., Sample adequately to estimate variograms of soil properties. *Journal of Soil Science*, vol. 43, 1992, p. 177-192.

S-GEMS: THE STANFORD GEOSTATISTICAL MODELING SOFTWARE: A TOOL FOR NEW ALGORITHMS DEVELOPMENT

NICOLAS REMY
Department of Geological & Environmental Sciences, Stanford University, CA 94305

Abstract. S-GeMS (Stanford Geostatistical Modeling Software) is a new cross-platform software for geostatistics. Capitalizing on the flexibility of the C++ Geostatistical Template Library (GsTL), it offers the more common geostatistics algorithms, such as kriging of one or more variables, sequential and multiple-point simulations. This software was developed with two aims in mind: be reasonably comprehensive and user-friendly, and serve as a development platform into which new algorithms can easily be integrated. S-GeMS is indeed built around a system of plug-ins which allow new geostatistical algorithms to be integrated, import/export filters to be added, new griding systems to be used such as unstructured grids.

The S-GeMS source code is made available to everyone to use and modify. It can be freely copied and redistributed.

1 Introduction

Geostatistics is an application-focused field, and as such requires the availability of flexible software. Yet most of the current geostatistical softwares lack user-friendliness, and their dated programming designs make it difficult to add new functionalities.

One of the main programming effort made publicly available is the *Geostatistical Software Library* (*GSLIB*) (Deutsch and Journel , 1998). *GSLIB* is a rich collection of geostatistics softwares, but as its authors themselves emphasize, it does not try to provide a user-friendly interface. *GSLIB* has served the geostatistics community well, but its now dated implementation design and awkward interface make it both impractical and an improbable development toolkit.

This spurred the development of a new cross-platform software: S-GeMS, the Stanford Geostatistical Modeling Software. S-GeMS retains most of the advantages of *GSLIB*:

- it is cross-platform: its source code is written in ANSI C++ and it only relies on libraries available to most operating systems. S-GeMS has been successfully compiled on Unix, Linux, Windows and Mac OSX.

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 865–871.

- It is possible to run S-GeMS algorithms in batch: hence tasks such as repetitive sensitivity analysis can be conveniently performed.
- The source code of the algorithms can be freely studied, modified and distributed. Based on modern programming language designs, many components of S-GeMS, most notably its geostatistics part, can be integrated into other software packages.

S-GeMS brings several additional features:

- a sleek graphical user interface, along with interactive 3D visualization
- improved input-output formats, while retaining backward-compatibility with the *GSLIB* data file format.
- a genuine development platform: S-GeMS was designed such that its capabilities could conveniently be augmented by the adjunct of plug-ins. These plug-ins can be used to add new geostatistics algorithms, support new file formats, or new grid data structures (e.g. faulted tetrahedralized solids).

2 Geostatistics in S-GeMS

2.1 AVAILABLE ALGORITHMS

S-GeMS provides many of the classical geostatistics algorithms, as well as some recent developments such as multiple-point statistics simulation. Among the geostatistics algorithms included in S-GeMS are the following:

Kriging The kriging algorithm in S-GeMS operates on variables defined on a constant volume support, in 3-D. It allows to account for non-stationarity of the mean $E(Z(\mathbf{u}))$ of the random function $Z(\mathbf{u})$ being estimated. That mean can either be known and constant: $E(Z(\mathbf{u})) = \mathrm{m}$ (simple kriging), unknown but locally constant (ordinary kriging), unknown following a trend: $E(Z(\mathbf{u})) = \sum_{\mathrm{k}} \mathrm{a_k f_k}(\mathbf{u})$, or known but locally varying. Goovaerts (1997) or Chilès and Delfiner (1999) for example, provide a detailed description of kriging and its variants.

Indicator kriging The indicator kriging algorithm can account for both hard equality-type data: $z(\mathbf{u}) = \mathrm{z_0}$ and inequality data: $z(\mathbf{u}) \in [\mathrm{a,b}]$ or $z(\mathbf{u}) < \mathrm{a}$ or $z(\mathbf{u}) > \mathrm{b}$

CoKriging A kriging which accounts for a secondary correlated variable. The means of both primary and secondary variables can be either known or unknown and locally constant.

Sequential Gaussian Simulation Although the theory of Gaussian simulation requires that each local conditional cumulative distribution function (ccdf) be estimated by simple kriging, the S-GeMS version of sequential Gaussian simulation allows for non-stationary behaviors by using other types of kriging such as ordinary kriging or kriging with a trend.

Sequential indicator simulation This algorithm can be used to simulate both continuous and categorical variables. Equality data, inequality data and secondary data can all be integrated into the simulated model.

Multiple-points statistics simulation Algorithms based on two-points statistics, e.g. all algorithms based on kriging, fail to model complex curvilinear heterogeneities. Images with very different structures (e.g. channels stretching from one end of the image to the other, and non-overlapping ellipses of "small" dimensions) can share the same two-point statistics (i.e. covariance). Two point-statistics are not enough to fully characterize complex (low-entropy) structures, and one must rely on higher order statistics (*multiple-point* statistics) to model these structures.

The snesim algorithm implements a sequential simulation algorithm, called single normal equation simulation (Strebelle, 2000), which simulates categorical variables using multiple-point statistics. In snesim, the multiple-point statistics are not provided by an analytical function but are directly borrowed from a training image. The training image depicts the type of structures and spatial patterns the simulated realizations should exhibit, and it needs not be constrained to any local data such as sampled values or well data. Training images can conveniently be generated by unconditional object-based simulations (Tjelmeland, 2000; Holden et al., 1998; Chilès and Delfiner, 1999).

Finally S-GeMS also provides elementary data analysis tools such as histograms, QQ-plots, scatter-plots, and interactive variogram modeling.

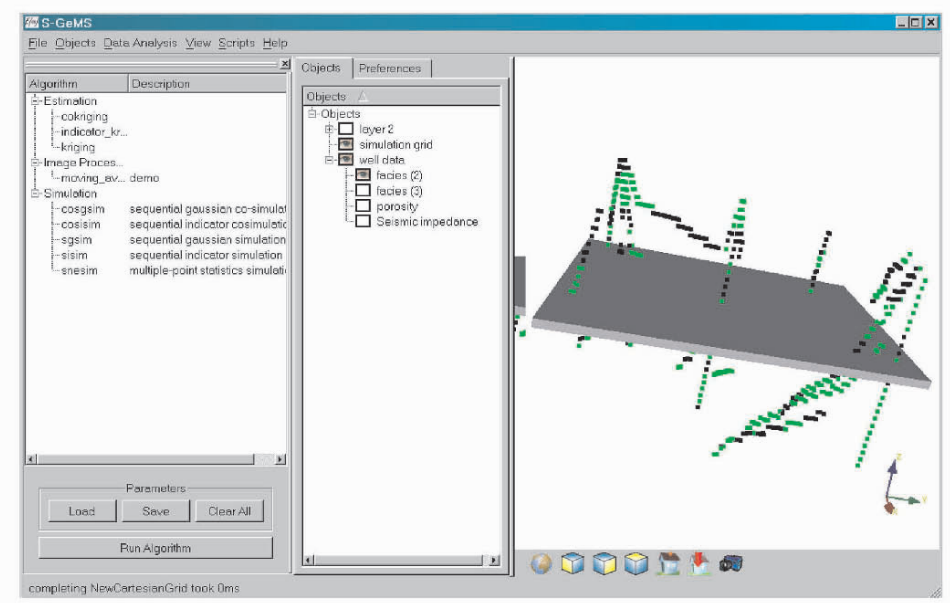
2.2 A SIMPLE S-GeMS SESSION

This section describes a simple S-GeMS session in which the multiple-point algorithm snesim is used to simulate a fluvial depositional environment, composed of a sand and shale facies, on a $100 * 130 * 30$ Cartesian grid. Facies data are available along several vertical and deviated wells. Fig. 1 shows the wells and a cross-section of the simulation grid. Cross-sections of the training image and display options such as colormap and cross-section selection are shown on Fig. 2.

The S-GeMS session proceeds as follows:

1. Load the data sets, i.e. the well data and the training image depicting the type of structures to be simulated.
2. Create a Cartesian grid with $100 * 130 * 30$ cells, call it *simulation grid*.
3. Select the snesim tool in the graphical interface and input the necessary parameters (see Fig. 2). In this example, 10 realizations are generated.
4. Display the result (i.e. the grid created in step 2), see Fig 3.

The parameters used for this run were saved into an XML file reproduced on Fig. 4.



Algorithm	Description
Estimation	
cokriging	
indicator_kr...	
kriging	
Image Proces...	
moving_av...	demo
Simulation	
cosgsim	sequential gaussian co-simulat
cosisim	sequential indicator cosimulatic
sgsim	sequential gaussian simulation
sisim	sequential indicator simulation
snesim	multiple-point statistics simulati

completing NewCartesianGrid took 0ms

Figure 1. S-GeMS main interface, with the well data and a horizontal cross–section of the simulation grid displayed. The middle panel is used to choose which object and object property to display

3 S-GeMS as a development platform

S-GeMS was designed to not only meet the practitioner's needs, but also provide a framework into which new ideas can conveniently be tested, and new algorithms easily implemented. S-GeMS provides two main facilities for the development of new algorithms: scripts written in the Python language (`http://www.python.org`), and plug-ins.

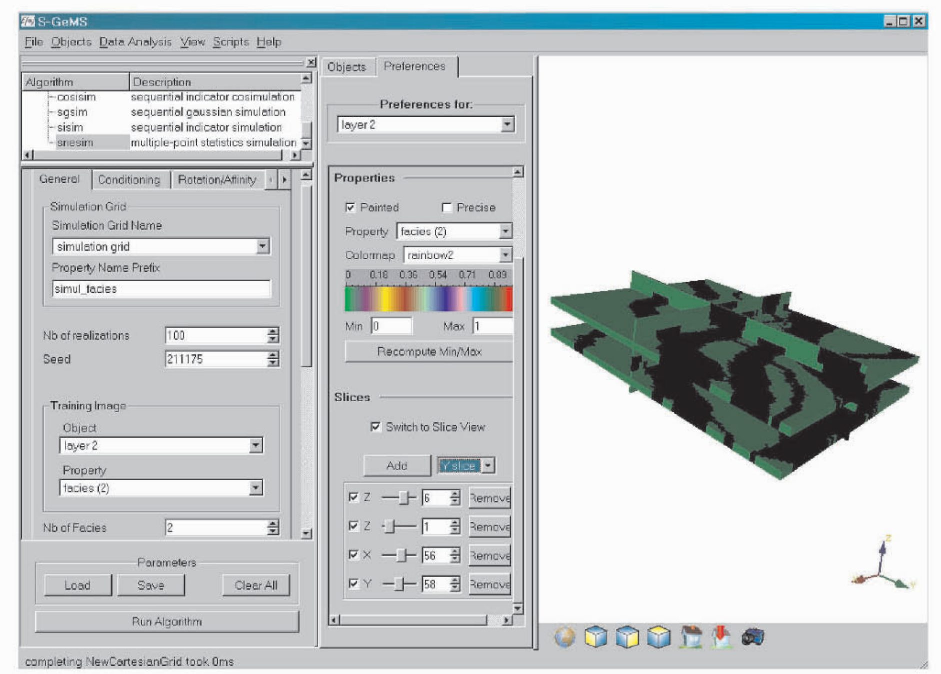
3.1 SCRIPTS

Most of the actions performed in S-GeMS with mouse clicks can also be executed by typing a command. These commands can be combined into a Python script to execute a complex sequence of operations. Python is a popular and powerful scripting language which supports most of the features of modern programming language, such as object-oriented programming, garbage collection or exceptions handling. Resources on Python can easily be found on the Internet.

Using scripts, it is for example possible to rapidly implement a cross-validation procedure to select the best variogram model.

3.2 PLUG-INS

New features can be added into S-GeMS through a system of plug-ins, i.e. pieces of software which can not be run by themselves but complement the main software. In



Algorithm	Description
cosisim	sequential indicator cosimulation
sgsim	sequential gaussian simulation
sisim	sequential indicator simulation
snesim	multiple-point statistics simulation

Simulation Grid

Simulation Grid Name: simulation grid

Property Name Prefix: simul_facies

Nb of realizations: 100

Seed: 211175

Training Image

Object: layer 2

Property: facies (2)

Nb of Facies: 2

Parameters

Preferences for: layer 2

Properties

☑ Painted ☐ Precise

Property: facies (2)

Colormap: rainbow2

Min: 0 Max: 1

Slices

☑ Switch to Slice View

Y slice

☑ Z 6

☑ Z 1

☑ X 56

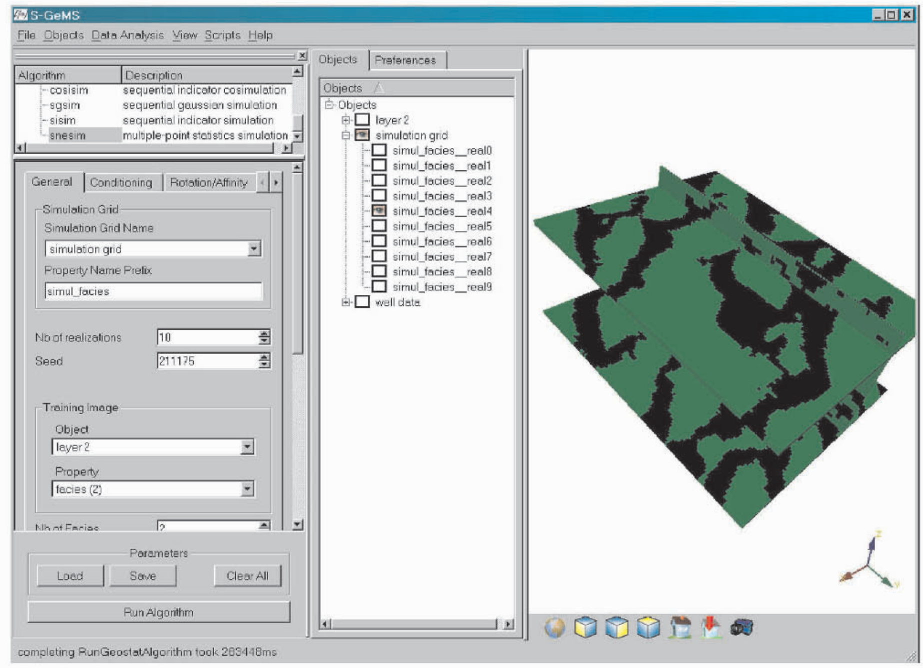
☑ Y 58

completing NewCartesianGrid took 0ms

Figure 2. The training image is displayed. The middle panel is used to select various display options, such as the colormap for a given property or which cross-section to display. Parameters for the snesim algorithm are input in the left panel

S-GeMS, plug-ins can be used to add new (geostatistics) tools, add new grid data structures, e.g. faulted stratigraphic grids, or define new import/export filters. It is compelling to implement new geostatistics algorithms as S-GeMS plug-in for several reasons:

- the plug-in can take advantage of the input/output facilities of S-GeMS. Geostatistics algorithms typically require several user-input parameters, for instance a data set and a variogram model. S-GeMS automatically generates a graphical interface for parameters input from a simple text description of the interface; hence no C++ code is required to gather the user's input. S-GeMS also provides several solutions for results output: results can be visualized in the 3D display of S-GeMS and saved into different file formats.
- The S-GeMS development toolkit includes several components generally needed by geostatistics algorithms such as data structures to represent data in space, whether organized on a Cartesian grid or unstructured (e.g. a set of unorganized points), or algorithms to search spatial neighbors on these data structures.
- All the components of the S-GeMS development toolkit are fully compatible with the *Geostatistics Template Library* (GsTL) concepts requirements


S-GeMS
File Objects Data Analysis View Scripts Help
General Conditioning Rotation/Affinity
Objects Preferences
Parameters
Load Save Clear All
Run Algorithm

Algorithm	Description
cosisim	sequential indicator cosimulation
sgsim	sequential gaussian simulation
sisim	sequential indicator simulation
snesim	multiple-point statistics simulation

Simulation Grid
Simulation Grid Name: simulation grid
Property Name Prefix: simul_facies

Nb of realizations: 10
Seed: 211175

Training Image
Object: layer 2
Property: facies (2)

Nb of Facies: 2

Objects
- layer 2
- simulation grid
  - simul_facies__real0
  - simul_facies__real1
  - simul_facies__real2
  - simul_facies__real3
  - simul_facies__real4
  - simul_facies__real5
  - simul_facies__real6
  - simul_facies__real7
  - simul_facies__real8
  - simul_facies__real9
- well data

completing RunGeostatAlgorithm took 283448ms

Figure 3. Object *simulation grid* now contains 10 new properties, one for each realization.

(Remy al., 2002), allowing the immediate use of the GsTL algorithms and data structures.

4 Conclusion

S-GeMS is a new geostatistics software freely available to the general public. To the end-user, it provides most of the classical geostatistics tools: data analysis, variogram modeling, kriging and simulation. User-friendliness of the software is achieved by a sleek graphical interface and an interactive 3D display of the data and results. S-GeMS capitalizes on the GsTL library to deliver algorithms performance often on par or higher than their (*GSLIB*) Fortran equivalents.

S-GeMS was also designed to be a platform for future developments. The S-GeMS API associated with the GsTL library provide an attractive framework in which to implement new algorithms. Owing to a system of plug-ins, these new algorithms can be conveniently integrated into S-GeMS. It is also possible to use scripts to perform complex tasks or quickly test new ideas within S-GeMS.

S-GeMS relies on four external libraries, all available under open-source licenses:

- Qt (`www.trolltech.com`) for the graphical interface
- Coin3D and SoQt (`www.coin3d.org`) for interactive 3D display

This page intentionally blank

```
<parameters>  <algorithm name="snesim" />
  <GridSelector_Sim  value="layer 2"  />
  <Property_Name_Sim  value="snesim_facies" />

  <Nb_Realizations  value="5" />
  <Seed  value="211175" />

  <PropertySelector_Training  grid="TI"   property="facies (2)"  />
  <Nb_Facies  value="2" />
  <Marginal_Cdf  value="0.7  0.3" />

  <Max_Cond  value="100" />
  <Search_Ellipsoid  value="10 10 3
                            0 0 0 " />

  <Hard_Data  grid="well data"   property="facies (2)"  />

  <Use_ProbField  value="0"  />
  <Use_Rotation  value="0"  />
  <Use_Affinity  value="0"  />

  <Cmin  value="1" />
  <Constraint_Marginal_ADVANCED  value="0.5" />
  <Nb_Multigrids_ADVANCED  value="4" />
  <Subgrid_choice  value="1"  />
  <Previously_simulated  value="4" />
</parameters>
```

Figure 4. SNESIM parameters

– GsTL (Remy, 2002) for the implementation of the geostatistics algorithms

The source code of S-GeMS is distributed under a Free Software license, meaning it can be freely copied, modified and redistributed. It is available on the Internet at
`http://pangea.stanford.edu/~nremy/SGeMS`

References

Chilès, J. and Delfiner, P., *Geostatistics: Modeling spatial uncertainty*, John Wiley & Sons, New York, 1999.

Deutsch, C.V. and Journel, A.G., *GSLIB: Geostatistical Software Library and User's Guide, 2nd Edition*, Oxford University Press, 1998.

Goovaerts, P., *Geostatistics for Natural Resources Evaluation*, Oxford University Press, 1997.

Holden, L., Hauge, R., Skare, O. and Skorstad, A., *Modeling of fluvial reservoirs with object models*, Mathematical Geology, vol.30, no. 5, p.473–496.

Remy, N., Schtuka, A., Levy, B. and Caers, J. *GsTL: the geostatistical template library in C++*, Computers & Geosciences, vol.28, p.971–979.

Strebelle, S. *Sequential simulation drawing structures from training images*, Ph.D. Thesis, Stanford University, 2000.

Tjelmeland, H. *Stochastic models in reservoir characterization and Markov random fields for compact objects* Ph.D. Thesis, Norwegian University of Science and Technology, 1996.

EVALUATING TECHNIQUES FOR MULTIVARIATE CLASSIFICATION OF NON-COLLOCATED SPATIAL DATA

SEAN A. MCKENNA
Geohydrology Department, Sandia National Laboratories
PO Box 5800 MS 0735, Albuquerque, New Mexico 87185 USA

Abstract. Multivariate spatial classification schemes such as regionalized classification or principal components analysis combined with kriging rely on all variables being collocated at the sample locations. In these approaches, classification of the multivariate data into a finite number of groups is done prior to the spatial estimation. However, in some cases, the variables may be sampled at different locations with the extreme case being complete heterotopy of the data set. In these situations, it is necessary to adapt existing techniques to work with non-collocated data. Two approaches are considered: 1) kriging of existing data onto a series of "collection points" where the classification into groups is completed and a measure of the degree of group membership is kriged to all other locations; and 2) independent kriging of all attributes to all locations after which the classification is done at each location.

Calculations are conducted using an existing groundwater chemistry data set in the upper Dakota aquifer in Kansas (USA) and previously examined using regionalized classification (Bohling, 1997). This data set has all variables measured at all locations. To test the ability of the first approach for dealing with non-collocated data, each variable is reestimated at each sample location through a cross-validation process and the reestimated values are then used in the regionalized classification. The second approach for non-collocated data requires independent kriging of each attribute across the entire domain prior to classification. Hierarchical and non-hierarchical classification of all vectors is completed and a computationally less burdensome classification approach, "sequential discrimination", is developed that constrains the classified vectors to be chosen from those with a minimal multivariate kriging variance. Resulting classification and uncertainty maps are compared between all non-collocated approaches as well as to the original collocated approach. The non-collocated approaches lead to significantly different group definitions compared to the collocated case. To some extent, these differences can be explained by the kriging variance of the estimated variables. Sequential discrimination of locations with a minimum multivariate kriging variance constraint produces slightly improved results relative to the collection point and the non-hierarchical classification of the estimated vectors.

1 Introduction

Examples of multivariate statistical techniques applied in the earth sciences include principal components analysis (PCA), Cluster Analysis (CA) and discriminant analysis (DA). A number of textbooks (e.g., Davis, 1988; Reyment and Savazzi, 1999) explain

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 873–882.

the basis of these techniques and their applications to earth science data sets. Applications of these multivariate techniques generally require a complete vector of the variables at each sampling location and a typical example problem is that of a geochemical survey where different chemical elements are assayed for each sample location. If some of the sample vectors are incomplete, several approaches have been devised to provide substitutes for the missing values. However, these approaches are for when incomplete sample vectors are the exception, not the rule. When the multivariate data exhibit complete heterotopy, a technique must be applied to construct the multivariate vectors prior to any classification or discrimination technique.

A host of techniques for regionalized, or spatial, classification of multivariate data have been developed. These include the use of a multivariate variogram (Bourgault and Marcotte, 1991) as a spatial weighting function (Bourgault et al., 1992) and/or combinations of principal components analysis coupled with factorial kriging (Goovaerts et al., 1993). However, in these approaches to multivariate spatial statistics, incomplete data vectors are also a problem. As pointed out by Wackernagel (1998), the multiple variable cross-variogram for entirely heterotopic data cannot be calculated and C(0) cannot be computed.

The problem of interest in this work is the classification of multivariate data sampled in a spatial domain into homogeneous groups. Kriging is used to investigate two approaches to dealing with the issue of complete data heterotopy: 1) Estimation of all variables at a series of "collection points" where the multivariate classification is done followed by estimation of class membership at all locations in the domain through kriging of either the probability of class membership or the generalized multivariate distance; and 2) Estimation of all variables at all locations in the spatial domain prior to classification. This first approach is similar to the regionalized classification method using full vectors as described by Bohling (1997) and Olea (1999). The two approaches are tested on a water chemistry data set (Bohling, 1997) and compared against the standard regionalized classification results obtained using the complete data vectors.

2 Classification and Discrimination

Classification is the process of assigning each member of a multivariate sample set to a finite number of *g* groups. This assignment is done on the basis of the same set of, *m*, variables contained in the vector, **x**, measured at each sample location. The distances between samples in *m*-dimensional multivariate space are calculated and the two samples with the shortest distance are combined. The position of this new group in *m*-dimensional space is now defined by the average of the combined variable values. This hierarchical grouping process is continued until a predefined number of groups remain.

The classification step is done using cluster analysis where the multivariate distance between any two vectors of data, $\mathbf{x}_r$ and $\mathbf{x}_s$, is calculated as the Mahalanobis distance:

$$d_{rs}^2 = (\mathbf{x}_r - \mathbf{x}_s)'\mathbf{C}^{-1}(\mathbf{x}_r - \mathbf{x}_s)$$

where **C** is the sample covariance matrix. The data vectors are clustered by using the calculated Mahalnobis distances with Ward's method:

$$d(r,s) = n_r n_s d_{rs}^2 / (n_r + n_s)$$

where n_r and n_s are the number of samples within clusters *r* and *s*, respectively. Ward's method begins with each vector representing a "cluster" in multivariate space and then it successively joins clusters so that the increase in the total within group sum of squares is minimized. The within group sum of squares for a cluster is defined as the sum of the squared distances between all points in the cluster and the centroid of that cluster. The hierarchical clustering process results in a dendrogram showing the successive grouping of individual vectors as the multivariate distance between clusters increases.

Discriminant analysis uses a set of previously classified samples as a training set to determine in which group any unclassified samples belong. Assignment of an unclassified sample to a group is done based on maximum posterior probability or minimum generalized squared distance. The latter approach is used in this work, where the generalized squared distance (after Bohling, 1997) between a sample and group *i* is:

$$D_i^2(\mathbf{x}) = d_i^2(\mathbf{x}) + \ln|C_i| - 2\ln(q_i)$$

where d^2 is the Mahalanobis distance and q_i is the prior probability of sampling from group *i*. This calculation of the generalized squared distance is developed within a Bayesian framework and under the assumption that multivariate normal density functions define the groups. A further simplification used herein is that the $\mathbf{C}_i$'s are equal across groups and therefore linear discriminant analysis can be used.

The *entropy of classification*, *H*, provides a convenient single measure of classification uncertainty across all posterior probabilities. (Bohling, 1997):

$$H = \frac{\left(-\sum_{k=1}^{g} p_k \ln p_k \right)}{\ln g}$$

where p_k is the posterior probability of membership in group *k* calculated using Bayes' theorem and assumed prior probabilities. As $p_k \to 1.0$ for any *k*, $H \to 0$. Conversely, *H* reaches a maximum value of 1.0 when all p_k are equal and the classification made was highly uncertain.

3 Regionalized Classification

When the spatial location and spatial covariance of the sample vectors are taken into account in multivariate classification and discrimination calculations, the procedure is referred to as "regionalized classification". Previous work in this area includes Harff and Davis (1990); Olea (1999) and Bohling (1997). The standard approach to regionalized classification uses all full vector sample locations to classify the data into groups. After groups are defined, discriminant analysis is then performed on the full vectors in a resubstitution (cross-validation) approach. This discrimination calculation provides the generalized squared distance from each vector to the centroid of each group identified in the classification procedure. The spatial covariance of these distances can be modelled and kriging is used to estimate these distances to all locations in the spatial domain. A separate covariance function must be identified for each group. For any location, group assignment is done by identifying the group to which the generalized squared distance is a minimum. This standard approach works well when full data vectors exist and serves as the baseline for comparison with two approaches below that do not rely on full vectors at the sample locations.

1.1 Collection Point

For heterotopic data, it may often be the case that one of the variables is oversampled relative to the others and these more numerous sample locations can serve as "collection points" onto which all other variables are estimated. Once this estimation is complete, these locations now serve as full vectors and the standard approach to regionalized classification can be conducted. A disadvantage of this approach is the required covariance inference for each of the *m* variables, with the exception of the oversampled variable that is not estimated, as well as for the *g* generalized squared distances.

1.2 Exhaustive Mapping

A conceptually simple alternative to the collection point method is to estimate all variables to all *n* unsampled locations across the spatial domain of interest and then do the classification at all spatial locations. This approach only requires spatial covariance inference for the *m* sampled variables as the classification procedure is only done after the spatial estimation. However, a major drawback of this approach is that hierarchical classification on large numbers of multivariate samples becomes a very memory intensive calculation as n(n-1)/2 distance comparisons must be calculated and stored. Non-hierarchical classification provides a less memory intensive alternative.

In keeping with the spirit of regionalized classification, it is recognized that classification is not simply a multivariate problem, but that there is also a spatial component, and, in this exhaustive mapping approach, the quality of the estimated variables varies spatially. *Sequential discrimination* is proposed here as a way to incorporate the quality of the estimated vectors directly into the assignment of samples to groups in the discriminant analysis step. Rather than classifying all locations at once, a classification proportion, *Pc*, of the estimated locations with the lowest overall multivariate kriging variance as summed across all variables are classified and then used as training data in a discriminant analysis to classify all other locations.

4 Example Calculations

1.3 Data Set

The example data set is a groundwater chemistry data set from the Dakota aquifer in Kansas and consists of six variables measured at each of 224 locations in a 450x325km region. This data set has been analysed previously (Bohling, 1997) and was obtained from the IAMG ftp site. The six variables are the log10 transform of the concentrations in mg/l of three cations (Ca, Mg, Na) and three anions (HCO3, SO4 and Cl). The univariate distributions are symmetric (mean/median ≈ 1.0) with coefficients of variation between 0.3 and 0.4. The univariate distribution of each variable is transformed to a standard normal distribution prior to classification or estimation.

1.4 Full Vector Classification

The full data vectors at each of the 224 locations are used with the standard regionalized classification approach to map group membership in the spatial domain. Following the approach of Bohling (1997), the clustering is stopped at four groups. These groups are arbitrarily numbered in terms of increasing Cl concentration and the spatial distribution of these groups at the sample locations are shown in Figure 1. Each full vector is also used in a discriminant analysis in resampling mode to determine the multivariate distance from each vector to the group to which it is assigned. These distances are then estimated at all unsampled locations and the final classification and the associated entropy of classification are shown as the top two images of Figure 2.

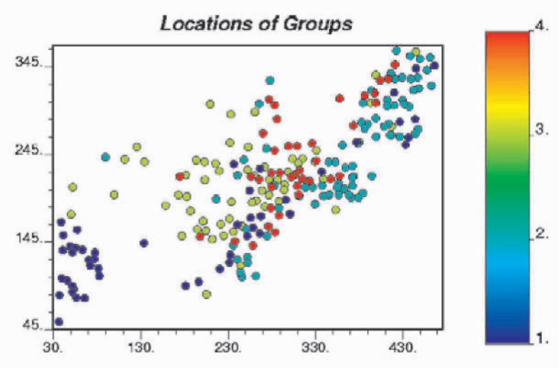


Figure 1. Locations of full vector sample data and groups defined from those full vectors. Distance units on axes are in km.

1.5 Collection Point Classification

The available data are isotopic, in order to make the data heterotopic, cross-validation, one sample left out at a time, is used to reestimate the values of the six variables at each of the 224 sample locations. In a practical application, only 5 (m-1) variables would be reestimated to the collection points with the actual measured values of the most densely sampled variable being used directly. These reestimated values are then considered as the full sample vectors and the standard regionalized classification approach is used as defined for the full vector case. The results of the classification and the associated uncertainty as defined by the entropy are shown in the middle two images of Figure 2.

1.6 Exhaustive Mapping Classification

Kriging is used to estimate values of each variable at all locations on a 5×5 km grid in the spatial domain. The spatial domain is small enough that hierarchical classification can be used to classify all vectors. The results of the exhaustive mapping classification are shown in the bottom images of Figure 2.

5 Results

Figure 2 shows that the three different approaches to regionalized classification produce three different results with the collection point approach producing the greatest level of mixing between the four groups. The standard, full vector, approach to regionalized classification is considered the baseline and the other two approaches are compared to it in Table 1. Table 1 shows the percent of the domain that was assigned to each group by each approach and it also contains the "mismatch sum" calculated as the sum of absolute differences between the groups identified by the standard (full vector) approach and each of the two other approaches as calculated across all locations.

Table 1. Results of three classification approaches.

Classification Approach	Groups				Mismatch Sum
	1	2	3	4	
Full Vector	19.2%	33.2%	33.9%	13.6%	NA
Collection Pt.	32.9%	37.3%	8.4%	21.3%	3210
Exhaustive	19.5%	33.4%	31.0%	16.0%	889

Table 1 shows that the percentage of the domain assigned to each group varies across the three approaches. Both the collection point and exhaustive mapping approaches give results that differ from the standard regionalized classification approach with the mismatch sum of the collection point approach being more than three times that of the exhaustive mapping approach.

The poor performance of the collection point approach is explained, to some extent, by the small number of locations, 224, used for the initial classification and the uncertainty in the variable values estimated at the collection points. Imprecision in the estimated values at the collection points leads to difficulty in consistently identifying the four groups. This result is highlighted by the entropy map for the collection point approach shown in Figure 2 (middle, right image) that shows values near 1.0 at most locations. These values of entropy indicate nearly equal probability of classification for each of the four groups.

6 Discussion

The results in Figure 2 and Table 1 show that the exhaustive mapping classification approach is superior to the collection point method. However, as the number of locations requiring classification becomes larger due to a larger domain or a finer spatial discretization, hierarchical classification methods will become computationally intractable. Two approaches to exhaustive classification on large fields are examined here: 1) initial classification of the *Pc* values with the lowest kriging variance followed by *sequential discrimination*. 2) k-means clustering, which is a non-hierarchical clustering technique. The steps involved in the sequential discrimination approach are:

1) Rank each estimated location in the spatial domain by summing the normalized kriging variances across all estimated variables. Because the number of samples and their spatial configuration is most likely different across the variables, this sum will vary considerably across the domain.

2) Select a proportion, *Pc*, of these locations that have the lowest summed variances and classify these locations into *g* groups. These will serve as the groups into which all remaining locations are assigned.
3) Select a proportion, *Pp*, of the remaining locations with the smallest summed variances and determine the multivariate distance from each sample within *Pp* to each group.
4) Assign the sample with the smallest multivariate distance within *Pp* to the closest group and add it to the set of training data.

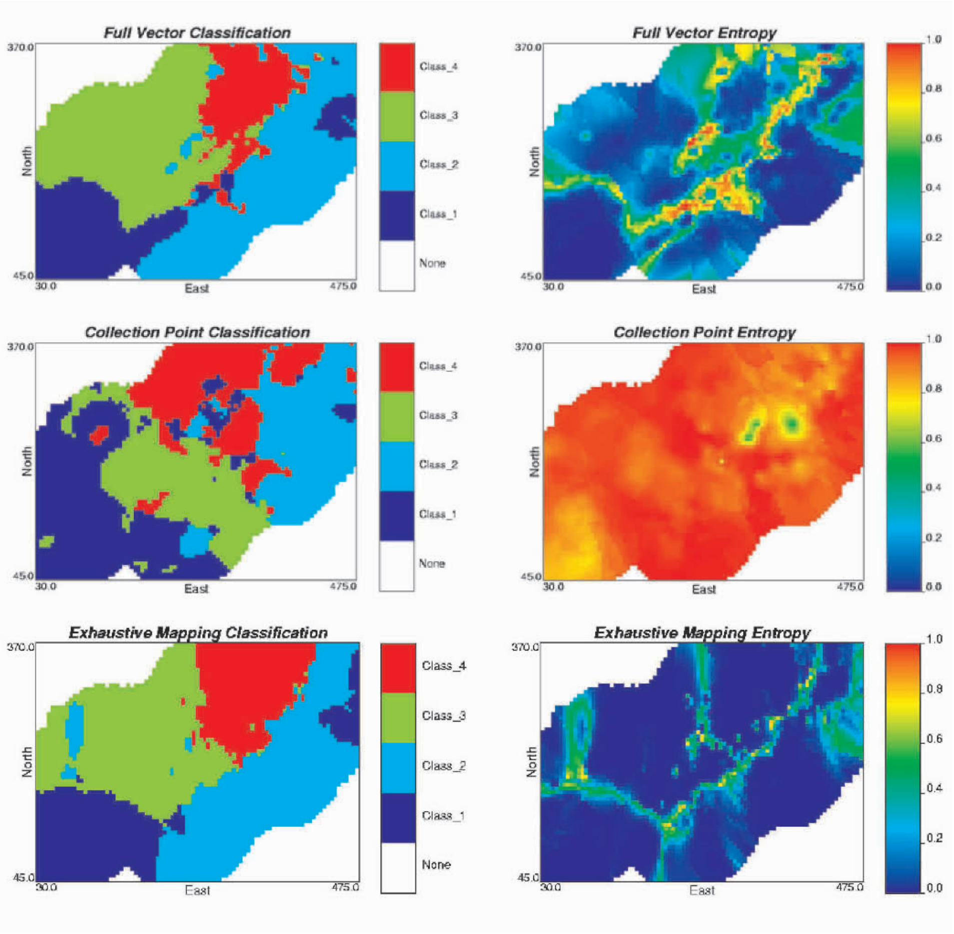


Figure 2. Maps of final group membership (left column) and entropy of classification (right column) for the full vector (top images), collection point (middle images) and exhaustive mapping approaches (bottom images).

From the list of the remaining unclassified samples, the one with the lowest variance sum is added to *Pp*, and the process is repeated starting at step 3 until all samples have been assigned. This sequential approach places a multivariate minimum kriging variance constraint on the assignment of locations to groups.

If *Pp* is set to be small enough to contain only a single sample, the unclassified sample with the smallest multivariate kriging variance will be assigned to an existing group. By keeping *Pp* large enough, the sample with the best fit to an existing group can be found from among a number of locations with relatively small kriging variances and this location will be the next one assigned.

Two different applications of the sequential approach were completed with *Pp* set to 0.05 and 0.20. *Pc* is set to 0.20 for all calculations. Sequential discrimination was done by iteratively selecting the location in *Pp* with the minimum value of the generalized squared distance and then adding the location with the next lowest sum of variances to *Pp*. The results of these two classifications and the results of the non-hierarchical classification are shown in Figure 3 and summarized in Table 2.

Table 2. Results of sequential classification approach on exhaustively mapped variables with different values of *Pp* compared to the full vector and non-hierarchical approaches.

Classification Approach	**Groups**				**Mismatch Sum**
	1	**2**	**3**	**4**	
Full Vector	19.2%	33.2%	33.9%	13.6%	NA
Exhaustive Pp = 0.05	28.1%	21.9%	30.8%	19.2%	1432
Exhaustive Pp = 0.20	27.9%	21.9%	31.0%	19.2%	1413
Non-Hierarchical: k-mean	15.1%	37.4%	29.8%	17.7%	2074

The choice of *Pp* makes a difference in the minimum generalized squared distances between the existing group centroids and the next sample to be classified. When *Pp* =0.05, the average minimum generalized squared distance across the first 3000 samples assigned to a group is 5.6. For the case of *Pp* = 0.20, the looser kriging variance constraint, the average minimum generalized squared distance decreases to 5.4. This smaller average minimum generalized squared distance leads to a slightly better match to the full vector results (Table 2).

Over all of the results, hierarchical classification of full vector data obtained at every location through kriging of each variable produce results that are most similar to case of isotopic data. The sequential discrimination approach developed here produces the results with the second lowest mismatch errors and these results are relatively insensitive to the choice of Pp. The sequential discrimination algorithm results are more similar to the full-vector results than are the non-hierarchical classification results.

It is noted that all of the classification approaches tested here that first require mapping of the variables to all locations through kriging lead to entropy of classification values that are low with respect to the entropy values calculated for the full vector approach. This result is an artifact of the estimation through kriging which acts as a smoothing operator thus reducing the calculated entropy values. Replacement of the kriging step with stochastic simulation of each variable followed by classification on each simulated field may allow for the calculation of more realistic entropy values. This stochastic simulation approach is reserved for future work.

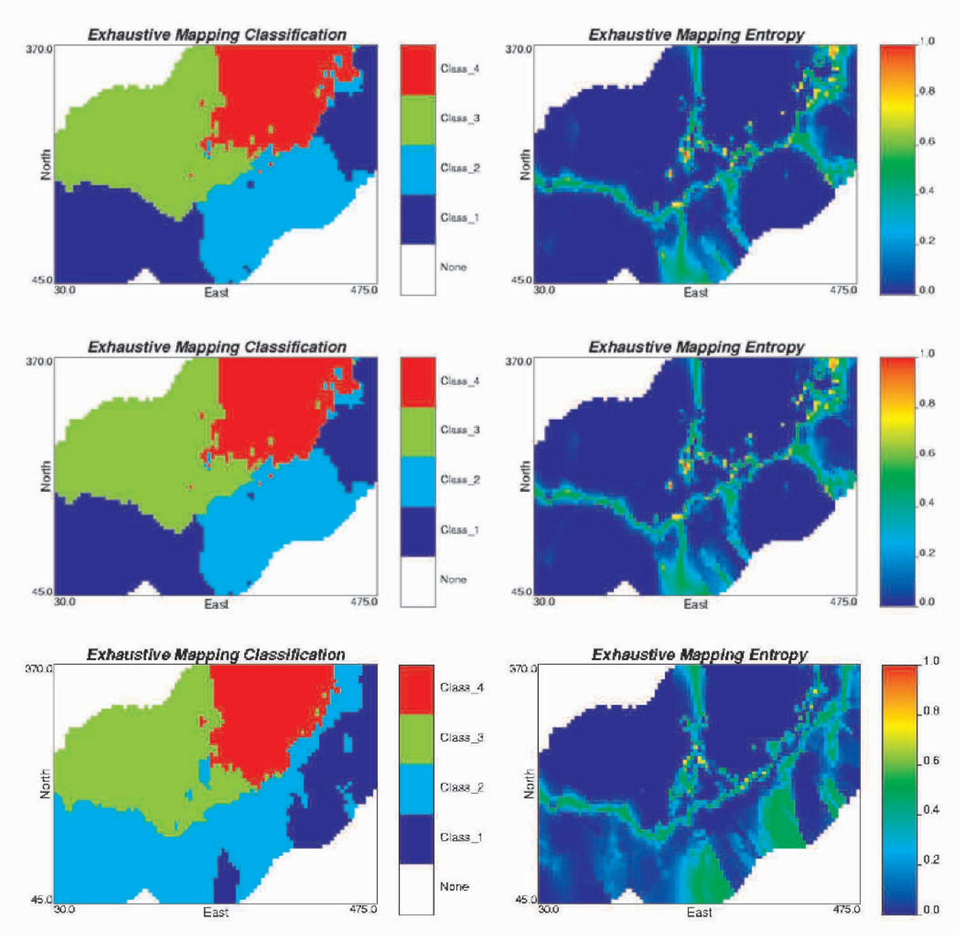


Figure 3. Classification results for the extended algorithm on the exhaustively mapped variables. The top images are results for *Pp* = 0.05, the middle images for *Pp* = 0.20 and the bottom images for the non-hierarchical classification. The left images show the classification and the right images show the entropy of classification.

7 Conclusions

Regionalized classification is an approach to determining spatially homogeneous groupings of data defined by multiple variables. While regionalized classification, as well as other multivariate mapping techniques have been developed for the case where all variables in the data set are collected at each location, little attention has been paid to the problem of regionalized classification from a completely heterotopic data set. This work examined two approaches to the scenario of heterotopic data: collection point and exhaustive mapping of the variables prior to classification. Additionally, a new algorithm was proposed for sequential assignment of estimated locations to previously defined groups under the constraint of those locations having a relatively low multivariate kriging variance. All approaches were tested on an example problem using a water chemistry data set. Results were compared to the standard regionalized classification approach completed with full data vectors at each sample location.

Results show that the collection point approach does not provide satisfactory regional classification. These results are most likely due to the relatively small number of collection points used in the initial multivariate classification and the necessity of two different sets of spatial estimation. The collection point approach resulted in nearly all groups having equal probability of occurrence at all locations as demonstrated by the high levels of classification entropy across the spatial domain.

The best results were obtained when all variables were estimated at all locations and then hierarchical classification was done across all locations. However, due to computational limits, this hierarchical classification approach will not be practical for large domains. The sequential discrimination approach where only a proportion of the estimated locations with the lowest multivariate kriging variances are used for the initial classification followed by sequential discrimination constrained to minimize the kriging variance of classified locations produces results that are less accurate than the hierarchical classification, but that are an improvement on the non-hierarchical classification approach. These results point out that the most important step in the regionalized classification procedure is the initial classification of the multivariate data into groups.

Acknowledgements

The author thanks Geoff Bohling for the data set and sharing his knowledge of regionalized classification. Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under contract DE-AC04-94-AL-85000.

References

Bohling, G.C., 1997, GSLIB-Style Programs for Discriminant Analysis and Regionalized Classification, *Computers and Geosciences*, 23 (7), pp. 739-761.

Bohling, G.C., J. Harff and J.C. Davis, 1990, Regionalized Classification: Ideas and Applications, in Proceedings of: Fifth Canadian/American Conference on Hydrogeology, Banff, Alberta, S. Bachu (ed.), Alberta Research Council, Edmonton, Alberta, Canada, pp. 229-242.

Bourgault, G., and D. Marcotte, 1991, Multivariable Variogram and its Application to the Linear Model of Coregionalization, *Mathematical Geology*, 23 (7), pp. 899-928.

Bourgault, G., D. Marcotte and P. Legendre, 1992, The Multivariate (Co)Variogram as a Spatial Weighting Function in Classification Methods, *Mathematical Geology*, 24 (5), pp. 463-478.

Davis, J.C., 1988, *Statistics and Data Analysis in Geology, 2nd Edition*, Wiley and Sons, New York.

Goovaerts, P, P. Sonnet and A. Navarre, 1993, Factorial Kriging Analysis of Springwater Contents in the Dyle River Basin, Belgium, *Water Resources Research*, 29 (7), pp. 2115-2125.

Harff, J. and J.C. Davis, 1990, Regionalization in Geology by Multivariate Classification, Mathematical Geology, 22(5), pp. 573-588.

Olea, R., 1999, *Geostatistics for Engineers and Earth Scientists*, Kluwer Academic Publishers, 303 pp.

Reyment, R.A. and E. Savazzi, 1999, *Aspects of Multivariate Statistical Analysis in Geology*, Elsevier, Amsterdam, 285 pp.

Wackernagel, H., 1998, *Multivariate Geostatistics, 2nd, Completely Revised Edition*, Springer, Berlin, 291 pp.

TRAVEL TIME SIMULATION OF RADIONUCLIDES IN A 200 M DEEP HETEROGENEOUS CLAY FORMATION LOCALLY DISTURBED BY EXCAVATION

MARIJKE HUYSMANS[1], ARNE BERCKMANS[2] AND
ALAIN DASSARGUES[1,3]

[1]Hydrogeology and Engineering Geology Group, Department of Geology-Geography, Katholieke Universiteit Leuven, Redingenstraat 16, 3000 Leuven, Belgium

[2]NIRAS/ONDRAF: Belgian Agency for Radioactive Waste and Enriched Fissile materials, Kunstlaan 14, 1210 Brussel, Belgium

[3]Hydrogeology, Department of Georesources, Geotechnologies and Building Materials, Université de Liège, Chemin des Chevreuils 1, Belgium

Abstract. In Belgium, the Boom Clay Formation at a depth of 200 m below surface is being evaluated as a potential host formation for the disposal of vitrified nuclear waste. The aim of this study is to model the transport of radionuclides through the clay, taking into account the geological heterogeneity and the excavation induced fractures around the galleries in which the waste will be stored. This is achieved by combining a transport model with geostatistical techniques used to simulate the geological heterogeneity and fractures of the host rock formation. Two different geostatistical methods to calculate the spatially variable hydraulic conductivity of the clay are compared. In the first approach, one dimensional direct sequential co-simulations of hydraulic conductivity are generated, using measurements of hydraulic conductivity (K) and 4 types of secondary variables: resistivity logs, gamma ray logs, grain size measurements and descriptions of the lithology, all measured in one borehole. In the second approach, three dimensional cokriging was performed, using hydraulic conductivity measurements, gamma ray and resistivity logs from the same borehole and a gamma ray log from a second borehole at a distance of approximately 2000 m from the first borehole. For both methods, simulations of the fractures around the excavation are generated based on information about the extent, orientation, spacing and aperture of excavation induced fractures, measured around similar underground galleries. Subsequently, the obtained 3D cokriged and 1D simulated values of hydraulic conductivity are each randomly combined with the simulated fractures and used as input for a transport model that calculates the transport by advection, diffusion, dispersion, adsorption and decay through the clay formation. This results in breakthrough curves of the radionuclide Tc-99 in the aquifers surrounding the Boom Clay that reflect the uncertainty of travel time through the clay. The breakthrough curves serve as a risk management tool in the evaluation of the suitability of the Boom Clay Formation as a host rock for vitrified nuclear waste storage. The results confirm previous calculations and increase confidence and robustness for future safety assessments.

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 883–892.

1 Aim of the study

The aim of this study is to model the transport of radionuclides through the Boom clay, taking into account the geological heterogeneity and the excavation induced fractures around the galleries in which the waste will be stored.

2 Geological context

The safe disposal of nuclear waste is an important environmental challenge. The Belgian nuclear repository program, conducted by ONDRAF/NIRAS, is in the process of characterizing the capacity of the Boom Clay as a natural barrier. At the nuclear zone of Mol/Dessel (province of Antwerp) an underground experimental facility (HADES-URF) was built in the Boom Clay at 225 m depth. In this area, the Boom Clay has a thickness of about 100 m and is overlain by approximately 180 m of water bearing sand formations. Several boreholes provide sets of geological, hydromechanical and geophysical data about the Boom clay.

The Boom clay is a marine sediment of Tertiary age (Rupelian) (Wouters and Vandenberghe 1994). The average hydraulic conductivity value of this formation is very low (K=2.10^{-12} m/s), but the clay is not completely homogeneous. It contains alternating horizontal sublayers of silt and clay with an average thickness of 0.50 m and a large lateral continuity (Vandenberghe et al. 1997). Furthermore, the clay exhibits excavation-induced fractures around the excavated galleries (Dehandschutter et al. 2002). The sublayers have hydraulic conductivity values up to 10^{-10} m/s (Wemaere et al. 2002) and the fractures may have even higher hydraulic conductivity values. These fractures are of a temporary nature as the clay has a considerable "self-healing" capacity.

3 Methodology

To test the robustness of earlier obtained results on the migration of radionuclides in the Boom clay two different approaches were used: a sequential simulation-based (Approach A) and a cokriging-based approach (Approach B). The main characteristics of the methods are summarized in Table 1.

	Approach A	**Approach B**
Dimensionality	1D	3D
Assumptions	horizontal continuity explicitly assumed	horizontal continuity not assumed
Hydraulic conductivity samples	52 K values from borehole A	42 K values from borehole A
Secondary data	- gamma ray of borehole A - resistivity of borehole A - 71 grain size measurements	- resistivity of borehole A - gamma ray of borehole A and B

	- lithology of borehole A	
Analysis and estimation / simulation of K	- data analysis - log transform of K - subdivision of Boom Clay into 3 zones - variogram and crossvariogram fitting - direct sequential simulations with histogram reproduction of hydraulic conductivity - 1D grid in z-direction with 0.2 m interval	- data analysis - no transform of K - no subdivision of Boom Clay - variogram and crossvariogram fitting - 3D cokriging of hydraulic conductivity - 8 by 85 nodes nodes in XYdir on a 25m grid, 1200 nodes in the Z dir on a 0.1m grid (612000 nodes in total)
Fracture simulation	Monte Carlo simulation from distributions based on observed fractures	Monte Carlo simulation from distributions based on observed fractures
Radionuclide	Tc-99	Tc-99
Position of radionuclide source	middle of Boom clay	middle of Boom clay
Hydrogeological model grid	A 20m by 15m by 102m box subdivided into 1m grid nodes in X, 0.17m grid nodes in the in Y and 0.2 to 1m grid nodes in the Z direction	A 100m by 125m by 106m box subdivided into 5m grid nodes in X and Y and 0.1m grid nodes in the Z direction.
Transport program	FRAC3DVS	FRAC3DVS

Table 1. Summary of the two methodologies

4 Data analysis

The value of incorporating the secondary information in the stochastic simulation of hydraulic conductivity was investigated by analyzing correlations between primary and secondary variables (Table 2). All secondary parameters show a fair to very good correlation with hydraulic conductivity and were therefore incorporated in the simulation of hydraulic conductivity.

	Correlation coefficient with hydraulic conductivity	
	Approach A	Approach B
Data	borehole A	borehole A and B
Electrical resistivity	0.73	0.87
Gamma ray	-0.65	-0.73
Grain size (d_{40})	0.95	N/A

Table 2. Correlation coefficients between measured hydraulic conductivity and secondary variables

5 Approach A: simulation of hydraulic conductivity on a 1D grid

5.1 VARIOGRAPHY

Previous geological work (Vandenberghe et al. 1997) on the Boom clay indicated that this formation can be subdivided in three units. For each unit, variograms and crossvariograms were calculated and modeled. The fitted models are given in Table 3. Figure 1a and 1b show two examples of experimental and fitted variograms and cross-variograms: the variogram of gamma ray of the Belsele-Waas Member and the cross-variogram of gamma ray and resitivity of the Belsele-Waas Member.

	Model	Nugget	Range	Sill
Boeretang Member	Spherical	0.035	4.6 m	0.03
Putte and Terhagen Member	Spherical	0.003	4.8 m	0.0056
Belsele-Waas	Spherical	0.23	5.5 m	0.38

Table 3. Fitted log K variograms of the three zones of the Boom Clay formation

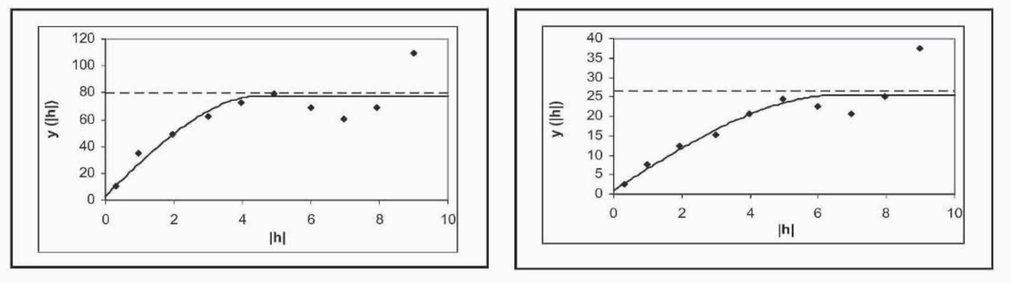


Figure 1. Experimental and fitted a) vertical variogram of gamma ray and b) vertical cross-variogram of gamma ray and resistivity of the Belsele-Waas Member.

5.2 SIMULATION

In this approach, the Boom clay is assumed to be laterally continuous. Therefore, one-dimensional vertical simulations of hydraulic conductivity were calculated on a dense grid in the Z direction (Fig. 2). These hydraulic conductivity values serve as input for the hydrogeological model. The simulation algorithm is iterative and contains the following steps:

1. The location to be simulated is randomly chosen. The spacing between the locations to be simulated was 0.2 m.
2. The simple co-kriging estimate and variance are calculated using the original primary and secondary data and all previously simulated values using COKB3D (Deutsch and Journel 1998).
3. The shape of the local distribution is determined in such a way that the original histogram of hydraulic conductivity is reproduced by the simulation. This is achieved by the following approach. Before the start of the simulation, a look-up table is constructed by generating non-standard Gaussian distributions by choosing regularly spaced mean values (approximately from -3.5 to 3.5) and variance values (approximately from 0 to 2).

The distribution of uncertainty in the data space can then be determined from back transformations of these non-standard univariate Gaussian distributions by back transformation of L regularly spaced quantiles, p^l, $l=1,...,L$:

$$K^l = F_K^{-1}\left[G\left(G^{-1}\left(p^l\right)\sigma_y + y^*\right)\right], \quad l = 1,..., L$$

where $F_K(K)$ is the cumulative distribution function from the original K variable, $G(y)$ is the standard normal cumulative distribution function, y^* and σ_y are the mean and standard deviation of the non-standard Gaussian distribution and the p^l, $l=1,...,L$ are uniformly distributed values between 0 and 1. From this look-up table the closest K-conditional distribution is retrieved by searching for the one with the closest mean and variance to the co-kriging values (Oz et al. 2003).

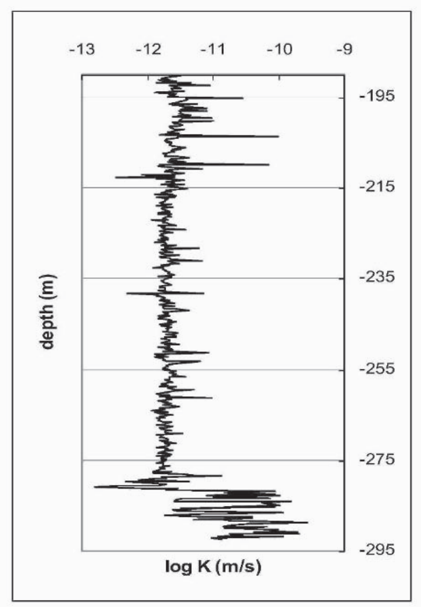


Figure 2. Simulation of the vertical hydraulic conductivity of the Boom Clay

4. A value is drawn from the K-conditional distribution by Monte-Carlo simulation and assigned to the location to be simulated. This approach creates realizations that reproduce (1) the local point and block data in the original data units, (2) the mean, variance and variogram of the variable and (3) the histogram of the variable (Oz et al. 2003).

6 Approach B: Estimation of hydraulic conductivity on a 3D grid

In the first approach a perfect horizontal layering was assumed and therefore the hydraulic conductivity values were simulated on a 1D vertical grid. In this approach we assume lateral variability and hydraulic conductivity is estimated on a 3D grid using information from boreholes 2000m apart.

6.1 VARIOGRAPHY

A coregionalization model was fitted to the variograms and cross-variograms. To model the horizontal continuity, gamma ray data are available in two boreholes 2048 m apart. The experimental horizontal variogram was calculated in 30 cm horizontal slices resulting in two average y(h) points, one at the origin and one at the interdistance between boreholes. The model in the horizontal direction has a range of 3 km. Fig. 3 illustrates the automatic sill fittings of the down hole cross-variograms of gamma ray and resistivity and resistivity and hydraulic conductivity.

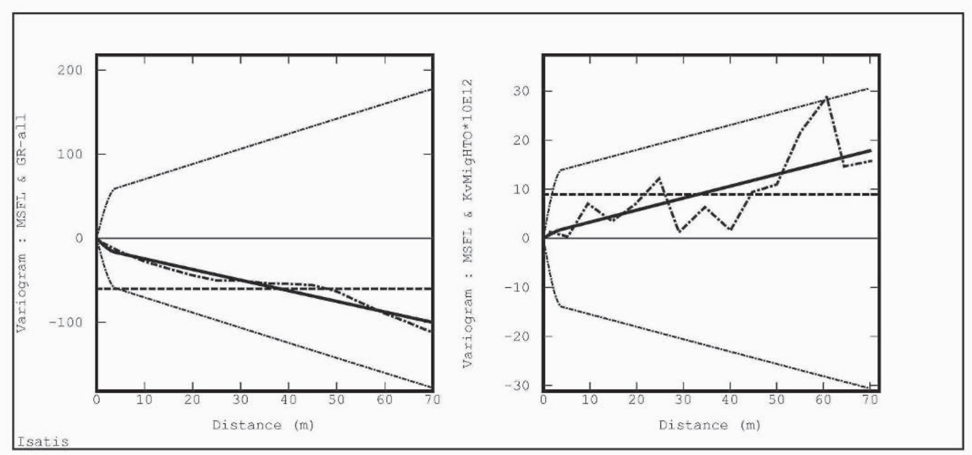


Figure 3. Cross-variograms of a) gamma ray and resistivity and b) resistivity and K.

6.2 ESTIMATION

The estimation is a straightforward cokriging on a 3D grid with a very fine mesh (10cm) in the Z direction. Fig. 4 illustrates one section between the two boreholes. There is more variability on the left side than on the right side of Fig. 4 due to the better conditioning of the co-kriging since the measured hydraulic conductivity values are only available in the borehole on the left.

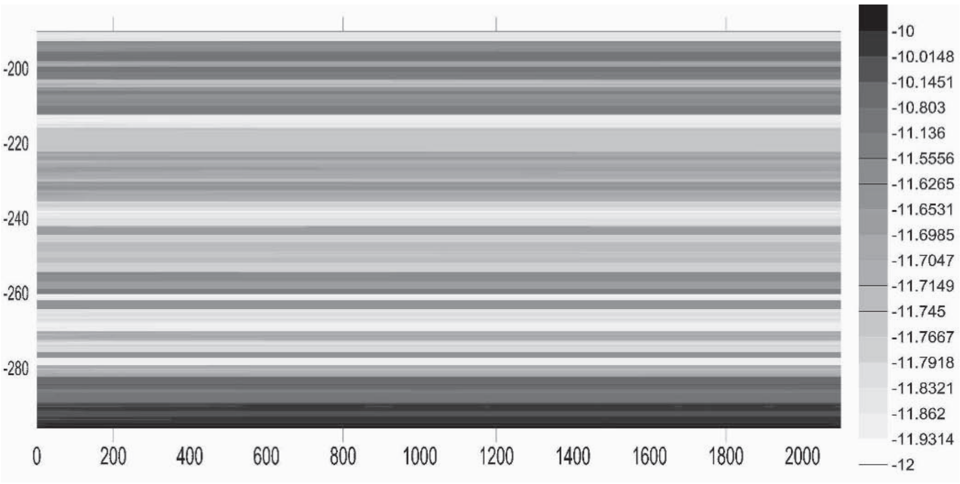


Figure 4. Logarithm of cokriged hydraulic conductivity in an YZ-profile

7 Simulation of fractures

Around the galleries in the Boom Clay, excavation-induced fractures are observed (Fig. 5). The excavation-induced fractures around the future disposal galleries were modeled as discrete fractures. Since these fractures will probably have similar properties to the fractures observed in previously excavated galleries in the Boom Clay, the input probability distributions of the fracture properties were derived from measurements carried out during recent tunnel excavation in the Boom Clay (Dehandschutter et al. 2002; Dehandschutter 2002; Mertens et al. 2004). These distributions are summarized in Table 4.

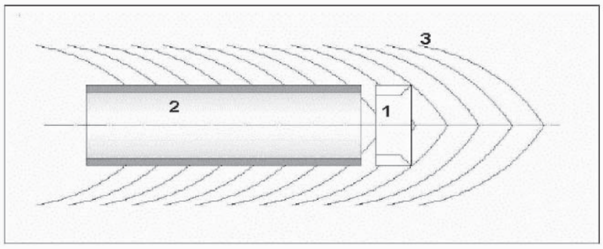


Figure 5. Schematic representation of a vertical cross section through the Connecting Gallery showing the typical symmetrical form of the encountered shear planes (1. Tunneling maching; 2. Supported tunnel; 3. Induced shear planes)

Variable	Distribution
fracture length	uniform distribution U (1m, 3m)
fracture aperture	uniform distribution U (0μm, 50μm)
fracture spacing	normal distribution $\mathcal{N}(0.7\text{m}, (0.12\text{m})^2)$
fracture dip	normal distribution $\mathcal{N}(53°, (11°)^2)$
fracture strike	perpendicular to the excavation

Table 4. Distributions of fractures properties around an excavated zone. The five distributions are assumed uncorrelated and independent.

8 Hydrogeological model

A local 3D hydrogeological model of the Boom Clay, including the estimation/simulations of matrix hydraulic conductivity values and the fractures, was constructed. The boundary conditions, radionuclide and source term model are the same for the hydrogeological models with the 1D K-simulations and the 3D co-kriged K-values. The model size and grid are different for both approaches (Table 1). The size of the model was a compromise between including as many fractures as possible and keeping the computation time manageable. The fine grid resolution in the Z direction was necessary to include the high resolution simulations of hydraulic conductivity and the geometry of the fractures. The vertical boundary conditions for groundwater flow are zero flux boundary conditions since the hydraulic gradient is vertical. Both approaches use the same Dirichlet horizontal boundary conditions. The specified head at the upper boundary is 2 m higher than the specified head at the lower boundary as the upward vertical hydraulic gradient is approximately 0.02 in the 100 m thick Boom Clay (Wemaere and Marivoet 1995). Zero concentration boundary conditions (Mallants et al. 1999) for transport are applied at the upper and lower boundaries of the clay since solutes reaching the boundaries are assumed to be flushed away immediately by advection in the overlying and underlying aquifers.

The model was calculated for the radionuclide Tc-99. Previous calculations revealed that this was one of the most important in terms of dose rates from a potential high-level waste repository for vitrified waste (Mallants et al. 1999). This radionuclide has a half-life of 213,000 years, a solubility limit of 3e-8 mole/l and a diffusion coefficient of 2e-10 m²/s and a diffusion accessible porosity of 0.30 was assumed for the clay. The transport processes considered in the model are advection, dispersion, molecular diffusion and radioactive decay.

The nuclear waste disposal galleries are assumed to be situated in the middle of the Boom Clay. The radionuclide source is modeled as a constant concentration source with a prescribed concentration equal to the solubility limit. The radionuclides slowly dissolve until exhaustion of the source.

For both approaches the radionuclide migration was calculated using FRAC3DVS, a simulator for three-dimensional groundwater flow and solute transport in porous, discretely-fractured porous or dual-porosity formations (Therrien et al. 1996, Therrien et al. 2003). The fractures were modeled as discrete planes with a saturated hydraulic conductivity of (Bear 1972):

$$K_f = \rho g \left(2b\right)^2 \Big/ \left(12\mu\right)$$

where ρ is the fluid density (kg/m³), g is the acceleration due to gravity (m/s²), $2b$ is the fracture aperture (m) and μ is the fluid viscosity (kg/(ms)). The model was run with the different simulations of hydraulic conductivity and fractures as input.

9 Results and discussion

9.1 RESULTS OF APPROACH A

Figure 6 shows the total Tc-99 fluxes through the lower clay-aquifer interface for 10 different simulations.

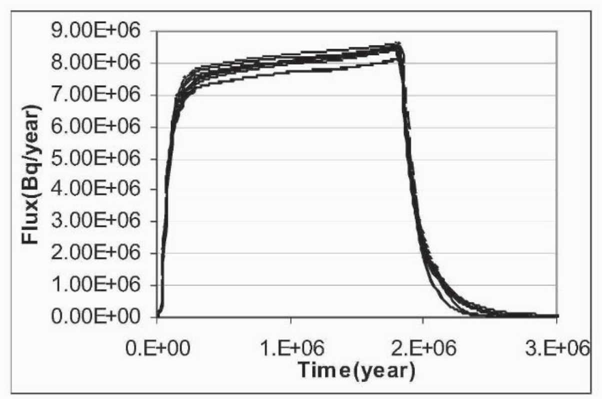


Figure 6. Total Tc-99 flux (Bq/year) through the lower clay-aquifer interface (1Bq = 1 disintegration / sec) calculated with approach A. Tc-99 has a half-life of 213,000 years.

The fluxes through the clay-aquifer interfaces increase relatively fast the first 200000 years. From 200,000 until 1,750,000 years, the fluxes increase more slowly. The fluxes decrease afterwards due to exhaustion of the source. The difference between the fluxes of the 10 different simulations is the largest in the time period from 200,000 till 1,750,000 years. The total amount of Tc-99 leaving the clay was calculated as the flux integrated over time for each simulation. The total Tc-99 masses leaving the clay vary between 1.423e+13 Bq and 1.541e+13 Bq through the lower clay-aquifer interface and between 1.443e+12 Bq and 1.489e+12 Bq through the upper clay-aquifer interface.

9.2 RESULTS OF APPROACH B

The Tc-99 fluxes calculated by this model were 1.45e13 Bq through the lower clay-aquifer interface and 1.38e13 Bq through the upper clay-aquifer interface (Fig.7). The flux through the lower clay-aquifer interface is in the range of fluxes calculated with the model with one-dimensional hydraulic conductivity simulations. The flux through the upper clay-aquifer interface is 4% smaller than the lowest flux calculated with approach A. These fluxes are thus approximately the same as the fluxes calculated by the model with 1-dimensional hydraulic conductivity simulations. This indicates that the assumption of perfect horizontal layering has no large effect on the calculated fluxes.

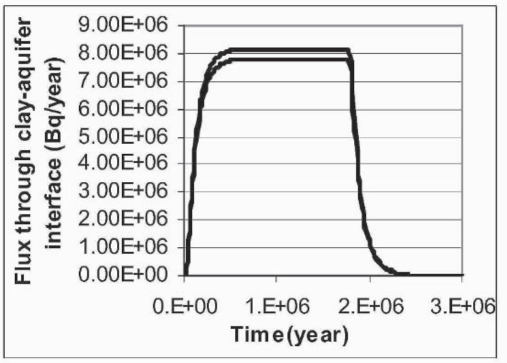


Figure 7. Total Tc-99 flux (Bq/year) through the lower and upper clay-aquifer interface calculated with approach B. Tc-99 has a half-life of 213,000 years.

9.3 DISCUSSION AND CONCLUSION

The range of total Tc-99 masses leaving the clay is rather small. The difference between the largest and the smallest calculated mass leaving the Boom clay is 8%. This result is important for the evaluation of the suitability of the Boom Clay Formation as a host rock for vitrified nuclear waste storage. The total mass fluxes leaving the clay, taking excavation induced fractures and high-conductivity sublayers into account, are not very different from the mass fluxes calculated by a simple homogeneous model. Changes in the modeled heterogeneity of hydraulic conductivity of the clay do not change the output fluxes significantly and therefore do not affect the ability of the clay to store vitrified nuclear waste in the predictive modeling. This again suggests that the Boom clay is a very robust barrier.

Acknowledgements

The Fund for Scientific Research – Flanders provided a Research Assistant scholarship to the first author. We thank ONDRAF/ NIRAS (Belgium agency for radioactive waste and enriched fissile materials) and SCK-CEN (Belgian Nuclear Research Centre) for providing data for this study. We thank René Therrien and Rob McLaren for providing Frac3dvs and for their assistance.

References

Allen D. et al., 1997, How to use Borehole Nuclear Magnetic Resonance. Oilfield Review Summer 1997, pp 34-57.

Bear J., 1972, *Dynamics of fluids in porous media*, American Elsevier, New York

Dehandschutter B., Sintubin M., Vandenberghe N., Vandycke S., Gaviglio P. and Wouters L., 2002, Fracture analysis in the Boom Clay (URF, Mol, Belgium), *Aardk. Mededel*., 12, 245-248

Dehandschutter B., 2002, Faulting and Fracturing during Connecting Gallery tunnelling at the URL at Mol (SCK-CEN), ONDRAF/NIRAS unpublished internal report

Dehandschutter B., Vandycke S., Sintubin M., Vandenberghe N., Gaviglio P., Sizun J.-P. and Wouters L.,2004, Microfabric of fractured Boom Clay at depth: a case study of brittle-ductile transitional clay behaviour, *Applied Clay Science*, in press

Deutsch C.V. and Journel A.G., 1998, *GSLIB geostatistical software library and user's guide*, Oxford University Press, New York

Mallants D., Sillen X. and Marivoet J., 1999, *Geological disposal of conditioned high-level and long lived radioactive waste: Consequence analysis of the disposal of vitrified high-level waste in the case of the normal evolution scenario,* NIROND report R-3383, Niras, Brussel

Mallants D., Marivoet J. and Sillen X., 2001, Performance assessment of vitrified high-level waste in a clay layer, *Journal of Nuclear Materials*, 298, 1-2, 125-135

Mertens J. and Wouters, L., 2003, 3D *Model of the Boom Clay around the HADES-URF*, NIROND report 2003-02, Niras, Brussel

Mertens J., Bastiaens W. and Dehandschutter B., 2004, Characterization of induced discontinuities in the Boom Clay around the underground excavations (URF, Mol, Belgium), *Applied Clay Science*, in press

Oz, B., Deutsch, C. V., Tran, T. T. and Xie, Y., 2003, DSSIM-HR: A FORTRAN 90 program for direct sequential simulation with histogram reproduction: *Computers & Geosciences*, v. 29, no.1, p. 39-51.

Vandenberghe N., Van Echelpoel E., Laenen B. and Lagrou D., 1997, *Cyclostratigraphy and climatic eustacy, example of the Rupelian stratotype*, Earth & Planetary Sciences, Academie des Sciences, Paris, vol. 321, p 305-315

Wemaere I. and Marivoet J., 1995, *Geological disposal of conditioned high-level and long lived radioactive waste: updated regional hydrogeological model for the Mol site (The north-eastern Belgium model) (R-3060)*, Niras, Brussel

Wemaere, I., Marivoet, J., Labat, S., Beaufays, R. and Maes, T., 2002, *Mol-1 borehole (April-May 1997): Core manipulations and determination of hydraulic conductivities in the laboratory (R-3590*)*,* Niras, Brussel

Wouters, L. and Vandenberghe, N., 1994, *Geologie van de Kempen: een synthese*: Niras, NIROND-94-11, Brussel

Therrien R. and Sudicky E.A., 1996, Three-dimensional analysis of variably-saturated flow and solute transport in discretely-fractured porous media, *Journal of Contaminant Hydrology*, 23, 1-2, p. 1-44.

Therrien R., Sudicky E.A. and McLaren R.G., 2003, *FRAC3DVS: An efficient simulator for three-dimensional, saturated-unsaturated groundwater flow and density dependent, chain-decay solute transport in porous, discretely-fractured porous or dual-porosity formations, User's guide*, 146 p.

GEOSTATISTICS AND SEQUENTIAL DATA ASSIMILATION

HANS WACKERNAGEL[1] and LAURENT BERTINO[2]
[1] *Centre de Géostatistique, Ecole des Mines de Paris, France*
[2] *Nansen Environmental and Remote Sensing Center, Norway*

Abstract. We review possibilities of introducing geostatistical concepts into the sequential assimilation of data into numerical models. The reduced rank square root filter and the ensemble Kalman filter are presented from this perspective. Contributions of geostatistics are discussed showing that sequential data assimilation is a promising area for the application of geostatistical techniques.

1 Introduction

Traditional geostatistical space-time geostatistics (Kyriakidis and Journel, 1999) is not able to take account of the generally strongly non-linear dynamics of multivariate space-time processes. To this effect physico-chemical transport models are in general more suitable. However, as the latter do not fully master the complexity of the processes they attempt to describe, either because of simplifying hypotheses or because the information serving to set up initial and boundary conditions is imperfect, it is appropriate to introduce statistical techniques in order to assimilate a flow of measurements emanating from automatic stations.

Recent projects at Centre de Géostatistique have permitted to explore these techniques in oceanography and air pollution. Soon it became evident that geostatistics could offer concepts and approaches to enhance Sequential Data Assimilation techniques. The thesis of Laurent Bertino (Bertino, 2001) and subsequent publications (Bertino et al., 2002; Bertino et al., 2003) have permitted to develop this theme.

More precisely, when dealing with Sequential Data Assimilation (as opposed to variational techniques) two viewpoints can be adopted. On one hand, from the point of view of the designer of deterministic numerical models, data assimilation is seen as an algorithm permitting to correct the state of the mechanistic model as new data comes in. On the other hand, from the point of view of the statistician the numerical model can help improve the operational prediction taking advantage of the knowledge of the non-linear relations between the various data sources.

It is this second viewpoint that we will privilege and seek to develop by positioning Geostatistics at the center of all data flows coming from a network of stations, a transport model coupled with other data sources or remote sensing data. Geostatis-

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 893–898.

tics, with its multivariate models, its anamorphosis and change-of-support models as well as its conditional simulation methods, offers a unique integrative framework for connecting these different informations, for understanding and modeling their statistical structure, for setting up prediction algorithms.

Data assimilation algorithms are needed for setting up operational forecasting systems as they are used in meteorology, oceanography, hydrology, ecology, epidemiology. An *operational forecasting system* consists of:

- a network of automatic stations,
- a dynamic forecasting model,
- a *data assimilation* algorithm.

As the station data generally provide only bad spatial coverage, the numerical model can compensate for this by a forecast based on known physical, chemical or biological relations. The data assimilation algorithm is then essential in combining these two sources of information sequentially in time, taking into account observational and model error.

This paper is divided into three sections. Section 2 describes a particular version of the extended Kalman filter, insisting on the geostatistical aspects. Section 3 presents another suboptimal Kalman filter which is close in spirit to the geostatistical simulation of Gaussian processes. Section 4 reviews a few possibilities of introducing geostatistical ideas into sequential data assimilation.

2 Kalman filter

We present the Kalman filter in its so-called *reduced rank square-root* (RRSQRT) version (Verlaan and Heemink, 1997) using notation that is close to the one used in geostatistics. Let $\mathbf{z}_t^o$ be the vector of the n observations at time t, $\mathbf{y}_t$ be the state of the system, we denote the state forecast with a f and the corrected state with a star, i.e.: $\mathbf{y}_t^f, \mathbf{y}_t^\star$. The forecast is performed by a numerical model $\mathcal{M}$, with boundary conditions $\mathbf{u}_t$, which describes the usually non-linear time dynamics. For computing error covariances we need to derive from the numerical model the tangent linear operator $\mathbf{M}$. We also need an observation linear operator $\mathbf{H}$ which serves both to transfer information from grid points to station locations and to generate from the forecast state "observations" as anticipated by the numerical model.

Leaving aside a detailed state space presentation of the Kalman filter, we merely present the algorithm which is composed of two steps. The first step is a propagation of the state from time $t-1$ to time t, using the numerical model to do the forecast:

$$\mathbf{y}_t^f \;=\; \mathcal{M}_t\left(\mathbf{y}_{t-1}^\star, \mathbf{u}_t\right) \tag{1}$$

and using the tangent linear operator to compute the corresponding error covariances,

$$\mathbf{C}_t^f \;=\; \mathbf{M}_t \mathbf{C}_{t-1}^\star \mathbf{M}_t^\top + \mathbf{Q}_t \tag{2}$$

The model noise covariance matrix $\mathbf{Q}_t$ needs to be carefully calibrated as it will condition the behavior of the filter. In a study of the hydrodynamics of the Baltic sea the sensitivity of the system stemmed mainly from the errors in the boundary conditions. Under the assumption that the water level field at the open boundary can be described by the wave equation a geostatistical model in the form of a space-time covariance model could be proposed (Wolf et al., 2001).

The second step is a correction of the state by kriging, perfomed at time t as soon as new data comes in. Kriging weights are computed from the forecast error covariances as well as the observation error covariances,

$$\mathbf{W}_t \;=\; \mathbf{C}_t^f \mathbf{H}^\top \left(\mathbf{H} \mathbf{C}_t^f \mathbf{H}^\top + \mathbf{C}^o \right)^{-1} \tag{3}$$

The corrected state is obtained by simple kriging

$$\mathbf{y}_t^\star \;=\; \mathbf{y}_t^f + \mathbf{W}_t \left(\mathbf{z}_t^o - \mathbf{H} \mathbf{y}_t^f \right) \tag{4}$$

and corresponding error covariances are computed,

$$\mathbf{C}_t^\star \;=\; \left(\mathbf{I} - \mathbf{W}_t \mathbf{H} \right) \mathbf{C}_t^f \tag{5}$$

In the RRSQRT algorithm the most important eigenvectors ("square roots") of the error covariance matrices $\mathbf{C}^f$, $\mathbf{C}^o$ are propagated ensuring both the positive definiteness of the matrices and a drastic reduction of dimensionality.

3 Ensemble Kalman filter

The *Ensemble Kalman* filter (EnKF) due to Geir Evensen (Evensen, 1994; Burgers et al., 1998) is based on a Monte-Carlo framework and has the avantage of not requiring a linearization of the numerical forecasting model $\mathcal{M}$. At each time step an ensemble of N forecasts

$$\left\{ \mathbf{y}_t^{f,i} = \mathcal{M}_{t-1}(\mathbf{y}_{t-1}^{\star,i}, \mathbf{q}_t^i);\; i = 1 \ldots N \right\}, \tag{6}$$

are propagated using simulated model errors $\{\mathbf{q}_0^i\}$. In geostatistical terms this first step can be seen as a *non-conditional simulation* generating N realisations of a non-stationary random function. The average forecast $\mathbf{y}_t^f$ and the covariance matrix $\mathbf{C}_t^f$ are computed directly on this ensemble of realizations.

The second step is the *conditioning* of the realizations by kriging on the basis of n observations collected at time t,

$$\left\{ \mathbf{y}_t^{\star,i} = \mathbf{y}_t^{f,i} + \mathbf{W}_t(\mathbf{z}_t - \mathcal{H} \mathbf{y}_t^{f,i} + \mathbf{u}_t^{o,i});\; i = 1 \ldots N \right\}, \tag{7}$$

where the observation errors are simulated according to a normal distribution $\mathcal{N}(0, \mathbf{C}^o)$ and the observation operator $\mathcal{H}$ is allowed to be non-linear. The first two moments of this ensemble of realizations approximate $\mathbf{y}_t^\star$ and $\mathbf{C}_t^\star$ in the same way as the mean of a number of conditional geostatistical simulations is equivalent to the solution of the kriging of the data. The details of the algorithmic implementation of the EnKF are discussed in (Evensen, 2003).

4 Contributions of geostatistics

We have seen that the correction step of the Kalman filter implies a kriging and that the EnKF is similar in spirit to the conditional simulations used in geostatistics. We also mentioned that geostatistics can be used to model the spatial correlation of the model error (Cañizares, 1999; Sénégas et al., 2001; Wolf et al., 2001).

UNIVERSALITY CONDITIONS

The correction step of the Kalman filter implies a simple kriging of the differences between the observations and their forecast according to the numerical model. It is possible to add universality conditions to this kriging in order to remove multiplicative or additive bias (Bertino, 2001). The approach is then equivalent to the one solved by external drift using numerical model output as external drift (Wackernagel et al., 2004). However, the difference is that geostatistics fits a covariance model to the forecast error at time t using some form of stationarity assumption, while in sequential data assimilation the covariances are propagated from the past and are not necessarily stationary.

In the RRSQRT filter the error covariance of the corrected state $\mathbf{C}_t^\star$ depends exclusively on the initial covariances $\mathbf{C}_0$, the model error covariances, the model operator, the location of observations through the matrix $\mathbf{H}$ and the observation error $\mathbf{C}_t^o$ (generally composed of white noise covariances). So the RRSQRT filter does not actually learn from the data but depends exclusively on how the error matrices were calibrated. The EnKF depends on the way how the errors $\mathbf{q}_t^i$ and $\mathbf{u}_t^{o,i}$ are generated, yet this affects only the mean and not the error covariances $\mathbf{C}_t^\star$.

The bias filtering through universality conditions in applications requires more stations than the five that were available in our study of the Odra lagoon. With a few stations only it turns out that the results may deteriorate when including universality conditions.

ANAMORPHOSIS OF NON-GAUSSIAN VARIABLES

The data assimilation methods presented above imply Gaussian assumptions. In applications the distributions may be skew and an anamorphosis, i.e. a transformation of the distribution, as used in non-linear geostatistics, may be of advantage. This idea was tested performing a lognormal transform to reduce skewness when implementing an EnKF for three variables (nutrients, phytoplancton, herbivores) in the context of a simplified ecological model of a water column in the ocean (Bertino et al., 2003). It turned out that data assimilation with anamorphosis generated smaller errors than with the standard EnKF. In particular, the spring bloom, which is the principal cause of non-linearity in the dynamics, less perturbs the filter with anamorphosis and the number of "false starts" of the phytoplancton bloom in springtime was significantly reduced.

MODELING THE SUPPORT EFFECT

It appears to be important in data assimilation problems to take account of the difference in the support of numerical model forecast and of observations, the support of the latter bein pointwise as compared to that of the state variables, which is of the size of the numerical model cells. Classical results in geostatistics have been adapted to the data assimilation context (Lajaunie and Wackernagel, 2000; Bertino, 2001).

For Gaussian state variables and observations the support correction resumes to an affine correction of the variances, because in the absence of bias the first moment of the observations is identical to that of the state variable by Cartier's relation.

In the framework of a lognormal model with assumption of permanence of lognormality for the different supports, merely the change of support coefficient needs to be inferred. The Gaussian anamorphosis generalizes the lognormal approach in the sense that it permits to transform an unspecified distribution towards a Gaussian distribution. The discrete Gaussian change of support model governed by a change of support coefficient can be applied in this context. Other change of support models like e.g. the gamma model studied by (Hu, 1988) could be used in the context of sequential data assimilation.

Finally it is also possible to work without an explicit change of support model in an approach based on geostatistical simulation on point support, where values on larger support are obtained by spatial averaging. By considering an ensemble of realizations empirical conditional distributions can then be easily computed.

The modelling of change of support has not yet been studied in detail and experimented in operational forecasting systems. This is due to the fact that there is a lack of awareness to implications of the support effect and this awareness is confined to domains in which geostatistics is already well known. Furthermore nonlinear geostatistical techniques need to be carefully adapted to applications in data assimilation in order to be able to add performance. A discrete Gaussian change of support model has been used for the purpose of downscaling air pollution forecasts at a resolution below that of the model cells by uniform conditioning (Wackernagel et al., 2004).

5 Conclusion

To keep the presentation simple, we have presented here the basic version of the ensemble Kalman filter as our main aim was to show the links and cross-fertilization potential between sequential data assimilation and geostatistical theory and methods. The EnKF is presently without doubt the most popular algorithm in sequential data assimilation (Mackenzie, 2003). Most recent developments (Evensen, 2004) can be found at the web site www.nersc.no/∼geir/EnKF/. Applications are found in many areas of operational forecasting for oceanography, meteorology, environmental and ecological monitoring. While the Kalman filter is a classical tool in hydrogeology (Eigbe et al., 1998), some new developments could occur in petroleum reservoir modelling (Naevdal et al., 2002).

References

Bertino, L. (2001). *Assimilation de Données pour la Prédiction de Paramètres Hydrodynamiques et Ecologiques: Cas de la Lagune de l'Oder.* Doctoral thesis, Ecole des Mines de Paris, Fontainebleau. http://pastel.paristech.org.

Bertino, L., Evensen, G., and Wackernagel, H. (2002). Combining geostatistics and Kalman filtering for data assimilation in an estuarine system. *Inverse Problems*, 18:1–23.

Bertino, L., Evensen, G., and Wackernagel, H. (2003). Sequential data assimilation techniques in oceanography. *International Statistical Review*, 71:223–241.

Burgers, G., van Leeuwen, P. J., and Evensen, G. (1998). On the analysis scheme in the ensemble Kalman filter. *Monthly Weather Review*, 126:1719–1724.

Cañizares, R. (1999). *On the Application of Data Assimilation in Regional Coastal Models.* PhD thesis, TU Delft, Rotterdam.

Eigbe, U., Beck, M. B., Wheather, H. S., and Hirano, F. (1998). Kalman filtering in groundwater flow modelling: problems and prospects. *Stochastic Hydrology and Hydraulics*, 12:15–32.

Evensen, G. (1994). Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte-Carlo methods to forecast error statistics. *J. Geophysical Research*, 99(C5):10143–10162.

Evensen, G. (2003). The Ensemble Kalman filter: Theoretical formulation and practical implementation. *Ocean Dynamics*, 53:343–367.

Evensen, G. (2004). Sampling strategies and square root analysis schemes for the EnKF. *Ocean Dynamics*, 54:539–560.

Hu, L. Y. (1988). *Mise en oeuvre du modèle gamma pour l'estimation des distributions spatiales.* Doctoral thesis, Ecole des Mines de Paris, Fontainebleau.

Kyriakidis, P. C. and Journel, A. G. (1999). Geostatistical space-time models: a review. *Mathematical Geology*, 31:651–684.

Lajaunie, C. and Wackernagel, H. (2000). Geostatistical approaches to change of support problems: Theoretical framework. IMPACT Project Deliverable Nr 19, Technical Report N–30/01/G, Centre de Géostatistique, Ecole des Mines de Paris, Fontainebleau.

Mackenzie, D. (2003). Ensemble Kalman filters bring weather models up to date. *SIAM News*, 36(8). http://www.siam.org/siamnews/10-03/tococt03.htm.

Naevdal, G., Mannseth, T., and Vefring, E. H. (2002). Near-well reservoir monitoring through ensemble Kalman filtering. *Society of Petroleum Engineers*, SPE 75235.

Sénégas, J., Wackernagel, H., Rosenthal, W., and Wolf, T. (2001). Error covariance modeling in sequential data assimilation. *Stochastic Environmental Research and Risk Assessment*, 15:65–86.

Verlaan, M. and Heemink, A. W. (1997). Tidal flow forecasting using reduced rank square root filters. *Stochastic Hydrology and Hydraulics*, 11(5):349–368.

Wackernagel, H., Lajaunie, C., Blond, N., Roth, C., and Vautard, R. (2004). Geostatistical risk mapping with chemical transport model output and ozone station data. *Ecological Modelling*, 179:177–185.

Wolf, T., Sénégas, J., Bertino, L., and Wackernagel, H. (2001). Application of data assimilation to three-dimensional hydrodynamics: the case of the Odra lagoon. In Monestiez, P., Allard, D., and Froidevaux, R., editors, *GeoENV II – Geostatistics for Environmental Applications*, pages 157–168, Amsterdam. Kluwer.

SPATIAL PROPERTIES OF SEASONAL RAINFALL IN SOUTHEAST ENGLAND

MARIE EKSTRÖM and ADRIAN CHAPPELL
Climatic Research Unit, University of East Anglia, UK
School of Environment & Life Sciences, University of Salford, UK

Abstract. Interpolated rainfall fields are important inputs to agricultural, hydrological and ecological models and serve to enhance our understanding of environmental systems and hence the natural and anthropogenic impacts on the surface environment. However, the large temporal and spatial variability in rainfall makes this variable difficult to estimate at unsampled locations, particularly when station network is highly dispersed. In this article, we present preliminary work within the Marie Curie Framework 6 project GAP (Geostatistical Analysis of Precipitation). Using one year of rainfall data, seasonal rainfall patterns are investigated for South East England. Both seasons showed anisotropic conditions with greater dependence in the W-E direction. The parameters of the models fitted to the directional variograms showed a larger range, but substantially smaller sill, for the summer season compared to those variograms of the winter season. Seasonal rainfall depth maps were derived using Simulated Annealing (SA) and Ordinary Kriging (OK). The two methods showed large differences in terms of local and regional variability, with smooth patterns associated with the OK maps and larger spatial variability associated with the SA maps. Both methods however, captured the large scale patterns that are typical to summer and winter rainfall in the study region. We suggest that the optimized patterns using SA could provide an alternative to OK, particularly for high temporal resolution rainfall data when OK produce unrealistically smooth maps.

1 Introduction

Spatial interpolation of rainfall at previously unsampled locations have proven a challenging task particularly when the rainfall network is sparse and with the presence of marked orography. A number of approaches have been used to estimate rainfall at unsampled locations. The simplest approach consisted of assigning to the unsampled location the record of the closest rain gauge (Thiessen, 1911) and it has been applied to the interpolation of point rainfall (Dirks *et al*., 1998). Mathematical functions (e.g., inverse-distance squared) are commonly used to weight the influence of surrounding values to the estimation of rainfall at unsampled locations (Bedient and Huber, 1992). Although these methods are quick and easy to use, they do not provide the user with information of spatial dependence nor an estimate of spatial uncertainty. Such information could however, be obtained using geostatistical methods. Furthermore, to aid the spatial estimation of rainfall patterns, geostatistical methods can incorporate secondary variables such as: orography, distance to sea and aspect (Agnew and Palutikof, 2000 and Prudhomme and Reed, 1999).

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 899–905.

In the Marie Curie Framework 6 project GAP (Geostatistical Analysis of Precipitation), the aim is to analyze spatial patterns of rainfall in the UK and the Iberian Peninsula using geostatistical tools. In this article we present a preliminary study to GAP, where seasonal patterns of rainfall in South East (SE) England are at focus. This region has been shown to be spatially coherent in terms of rainfall (Wigley *et al*., 1984) and comprises a sufficient number and density of rainfall gauges for geostatistical analysis.

In the first part of the analysis, we produced directional and anisotropic variograms for seasonal rainfall totals. The variograms were fitted with authorised models and their parameters used to interpret the spatial characteristics of the rainfall. In the second part, maps of seasonal rainfall depth were created using Ordinary Kriging (OK), and Simulated Annealing (SA). The OK method is already an established method to estimate rainfall, although it is usually not preferred in combination with rainfall data of high temporal resolution, as it creates too a smooth map. For this reason, we chose to test an alternative and much less conventional method, the SA. Unlike the OK, which is effectively an optimised interpolator (Armstrong, 1998), the SA is an optimising method that can generate alternate conditional stochastic images of either continuous or categorical variables (Deutsch and Journel, 1998). The stochastic component of the method allows the creation of maps with larger local and regional variability. There are few examples in the literature of SA being used within a climatological framework (one example being Pardo-Igúzquiza, 1998), and further research is needed to investigate the applicability of SA to mapping of rainfall data.

2 Rainfall data

Daily rainfall totals were provided from the UK Met Office database (MIDAS) via the British Atmospheric Data Centre (http://www.badc.rl.ac.uk). An inventory of available records for the period 1960–2000 showed that the largest number of available station records existed for the year 1972 (443 records). Of these, six records contained some degree of missing data; hence 437 records remained for the data analysis.

For each station, seasonal rainfall depths/or rainfall totals (mm) were computed for the meteorological seasons winter: December, January and February, and summer: June to August — the autumn and spring seasons are not presented here as we want to show the maximum difference in rainfall patterns with respect to rainfall genesis. A seasonal partitioning of the data is preferable as different rainfall processes are dominant during the different seasons. In summer, SE England experiences both convective and frontal rainfall whilst during winter frontal rainfall is dominant.

Plots of the seasonal totals as circles, scaled according to their relative magnitude, gives an initial image of the spatial differences in rainfall depth between summer (Figure 1a) and winter (Figure 1b). During summer, the maximum rainfall depth was 206 mm and the minimum 48 mm (Figure 1a). Larger rainfall depths are found at the eastern most tip of the study region, which may reflect the influence of convective storms originating from the continent (Jones and Read, 2001). During the winter of 1972, the largest and smallest amount of station rainfall was 396 mm and 119 mm respectively (Figure 1b). A decrease of rainfall from west to east is evident in the data, indicating the main path of frontal depressions embedded within the mid-latitude westwind belt.

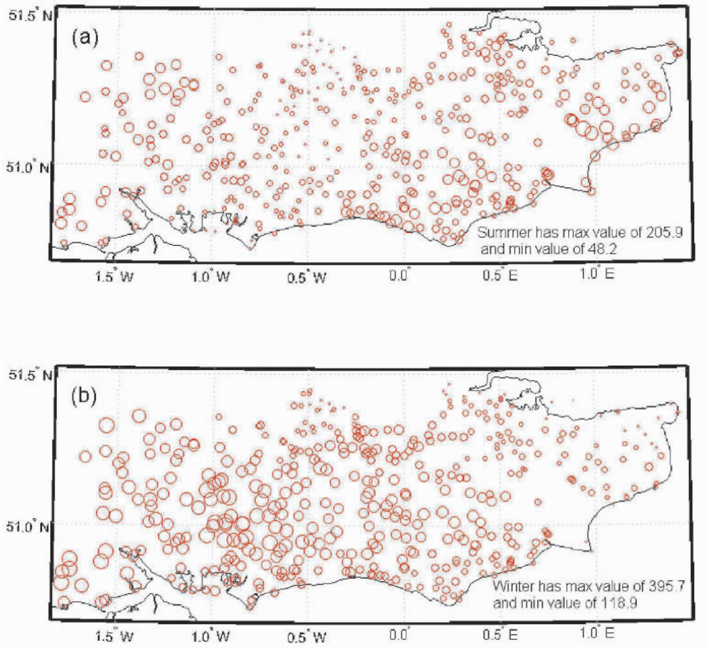


Figure 1. Rainfall depth (mm) for (a) summer (June to August) and (b) winter (December to February). The size of the circles is scaled relative to the rainfall depth, where the largest and smallest sized circles refer to the maximum and minimum rainfall depth respectively.

3 Choice of theoretical variogram

Sample variograms were calculated for four directions (N-S, NE-SW, W-E and NW-SE, 22 degrees tolerance). Each of these showed the presence of trend in both seasons (not shown). An attempt to remove the trend was made by fitting a low-order polynomial to the data. Using the residuals, a new set of directional sample variograms were calculated. These were fitted to separate authorised variogram models (spherical, exponential, circular, Gaussian and power) using weighted least squares in Genstat (Genstat 5 Committee, 1992). Selection of the best fitted model was based on the smallest mean squared error (MSE) and the results are shown in Table 1.

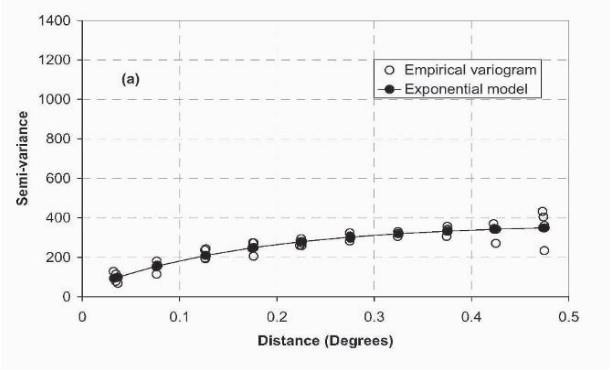
In both seasons, the greatest spatial dependence (i.e., the *range*), was found in the W-E direction; 1.56 deg (~170 km) in summer and 0.40 deg (~45 km) in winter. For most directions, the range was shorter for the winter season compared to the summer season. The range of residual rainfall in summer in all other directions was ~0.32 deg (~35 km), and during winter it was 0.29 deg (~ 30 km) in the NW-SE direction and ~0.17 deg (~ 20 km) in the NE-SW and N-S directions. An overall difference between the seasons was also evident in the variance (i.e., the *sill*), as shown by the substantially larger values during winter compared to summer. In summer, the largest variation was found along the W-E direction, followed by the NW-SE, N-S, and NE-SW direction. In winter, the largest variation was found in the NW-SE direction followed by the N-S, W-E, and NE-SW direction. The nugget is the proportion of unexplained variance, hence it may be useful to look at the proportion of the nugget to the total sill. We refer to it as *nugget-to-sill ratio*. During summer, the largest nugget-to-sill ratio is found in the N-S and NW-SE directions, whilst in winter the largest nugget-to-sill ratio is given in the N-S and NE-SW directions.

Season	**Direction**	**Model**	**Range**	**Sill**	**Nugget**	**$c_0/(c+c_0)$**
			(a)	**(c)**	**(c_0)**	**(%)**
Summer	N-S	Spherical	0.30	204.3	98.3	32.5
	NE-SW	Spherical	0.32	255.6	0.8	0.3
	W-E	Exponential	0.52*	551.2	57.5	9.4
	NW-SE	Spherical	0.35	227.7	83.5	26.8
	Anisotropic	Exponential	0.17*	335.4	35.5	9.6
Winter	N-S	Spherical	0.19	774.0	171.2	18.1
	NE-SW	Spherical	0.14	625.4	111.0	15.1
	W-E	Circular	0.40	870.0	32.7	3.6
	NW-SE	Circular	0.29	1006.0	108.0	9.7
	Anisotropic	Spherical	0.31	656.3	232.9	26.2

*The effective range for the exponential variogram is 3a.

Table 1. Values of the model parameters fitted to variograms for summer and winter residual rainfall from a fitted polynomial. Unit of range is degrees and *c* is the spatially dependent sill.

Because the directional variograms showed clear differences in range and sill, it was not appropriate to assume isotropic conditions in the study area. Hence, anisotropic variograms were calculated for each season using weighted least squares in Genstat (Genstat 5 Committee, 1992) (Table 1 and Figure 2). The fitted anisotropic model of summer residual rainfall had a larger range 0.2 deg (~20 km), but smaller spatially dependent sill, than that of the fitted anisotropic model of the winter residual rainfall. The nugget-to-sill ratio for the model fitted to winter data was 26 % (compared to 9.6 % in summer) indicating that a large proportion of the variation was not explained by the fitted model.



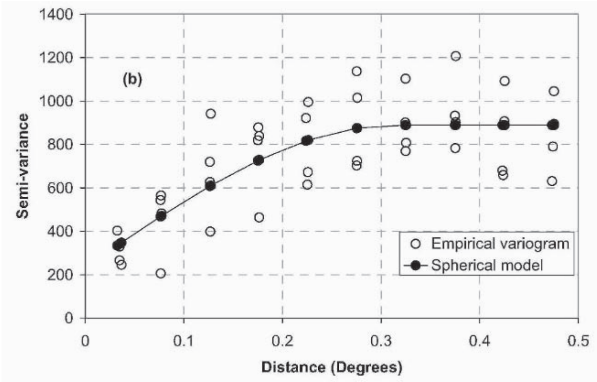


Figure 2. Anisotropic variograms for (a) summer and (b) winter. Lag distance is 0.05, see Table 1 for parameters of the variograms.

4 Estimation of gridded residuals

Two methods were used to produce high resolution (0.02 latitude by 0.02 longitude) grids for SE England: Ordinary Kriging (OK) (Figure 3) and Simulated Annealing (SA) (Figure 4). The OK estimates a grid by minimising the MSE between the estimated and the true field, using the parameters of the anisotropic variogram to describe the spatial structure of the residual rainfall. In this article the GSLIB function KT3D (Deutsch and Journel, 1998) was used to derive the kriging estimates. The SA is a stochastic global minimisation technique that mimics the metallurgical process of annealing. An imaginary analogy is made between the slowly cooling metal and the optimisation of the rainfall grid. At an initial stage of high temperatures, the molecules of the molten metal move relatively freely and reorder themselves into a very low energy structure. In SA this process is recreated by allowing gridded values swap places. A perturbation of the grid is accepted if an objective function is lowered (Deutsch and Journel, 1998). However, the acceptation of a perturbation is not only dependent on the objective function. The higher the "temperature" the greater the probability that an "unfavourable" perturbation is accepted (Deutsch and Journel, 1998). In this article, the SA was carried out using the GSLIB function SASIM (Deutsch and Journel, 1998), where the SA was constrained by a smoothed histogram of the rainfall depth residuals and the anisotropic variogram. For further information on SA please see: Deutsch and Journel (1998), Chilès and Delfiner (1999), and Lantuéjoul (2002).

5 Results

The gridded residuals from both the OK and SA were added back to the low-order polynomial to give actual rainfall depth (Figure 3 and 4). Altough the two sets of maps have rather different appearance, they share the same large scale features. The magnitude of rainfall is altogether smaller during summer. Local increases in rainfall are controlled by orography and, along the southeast coast, are influenced by convective storms encroaching from the continent. The winter pattern shows larger variability with increasing rainfall depths in the SW.

In comparison to the SA maps, the OK maps display smoothed surfaces and much less variability for both summer and winter seasons. Some smoothness may be expected as the data is in the form of seasonal totals. For example, larger variability would have been expected if daily, or even multi-day totals, had been used. Nevertheless, in this study, a large proportion of the smoothness in the OK maps is due to spatial interpolation. SA, on the other hand, provides larger variability whilst still honouring the characteristics of the data in the histogram and the spatial structure of the variogram.

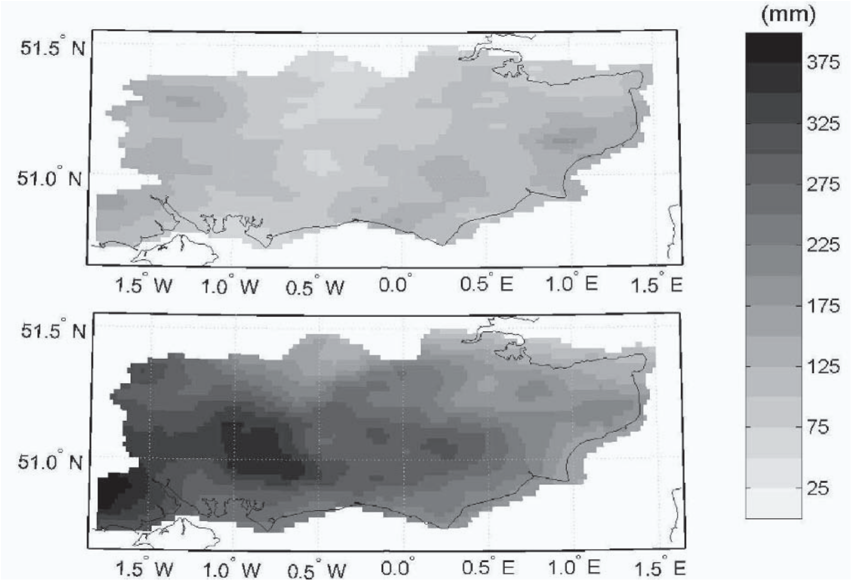


Figure 3. SE England rainfall depth during (a) summer and (b) winter, during one year (1972), as estimated using ordinary kriging.

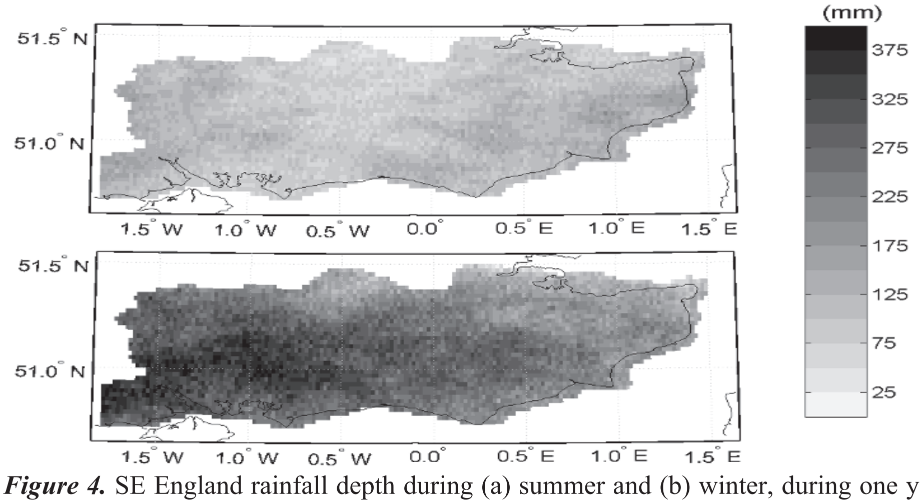


Figure 4. SE England rainfall depth during (a) summer and (b) winter, during one year (1972), as estimated using simulated annealing.

This page intentionally blank

6 Discussion and Conclusions

The seasonal rainfall depth for SE England was analyzed for one year. Directional variograms revealed anisotropic conditions in the data, with the strongest spatial dependence in the W-E direction for both seasons. Overall, the summer variograms had a larger range but smaller sill than the winter variograms. We did, however, expect a larger range during winter compared to summer. The reason being that winter rainfall is usually associated with frontal depressions whilst summer rainfall is usually associated with the smaller scaled convective processes. The smaller range during winter in this study is probably due to the study area being too small to capture the real scale of rainfall processes.

Two sets of gridded maps were produced using OK and SA. Whilst both methods showed the same overall spatial features, the OK maps showed stronger smoothing of the data compared to the SA maps. The maps of rainfall produced using SA gave a more reasonable representation of the variability in rainfall depth compared to that of the smoothed OK map. Further analysis will investigate the use of SA to estimate spatial patterns of rainfall. Issues such as: how to best represent multiple SA generated patters, what is the correct level of fluctuation in the SA patterns, how sensitive is the method to the rate of the “cooling” etc. all needs to be better understood. Nevertheless, we believe that the SA could prove a valuable method to investigate variability and uncertainty in rainfall mapping.

Acknowledgements

This work was funded by a Framework 6 Marie-Curie Intra-European Fellowship (n° MEIF-CT-2003-501494). The authors would also like to thank the two anonymous reviewers whose comments helped to improve the paper.

References

Agnew, M.D. and Palutikof, J.P., GIS-based construction of baseline climatologies for the Mediterranean using terrain variables, *Climate Research*, vol. 14, 2000, p.115-127.

Armstrong, M., *Basic Linear Geostatistics*. Springer-Verlag, Berlin Heidelberg, 1998.

Bedient, P.B. and Huber, W.C., *Hydrology and Floodplain Analysis*. 2nd edition, Addison-Wesley, Reading, MA, 1992.

Chilès J.-P. and Delfiner, P., *Geostatistics- Modeling Spatial Uncertainty*. Wiley series in probability and statistics, John Wiley & Sons, inc, USA, 1999.

Deutsch C.V., Journel A.G., *GSLIB Geostatistical Software Library and User's Guide*. 2nd Ed. Oxford University Press, 1998.

Dirks, K. N., Hay, J.E., J, E., Stow, C.D. and Harris, D., High-resolution studies of rainfall on Norfolk Island Pert II: interpolation of rainfall data. *Journal of Hydrology*, vol. 208, 1998, p.187-193.

Genstat 5 Committee., *Genstat 5, Release 3, Reference Manual,* Oxford University Press, Oxford, 1992.

Jones, P.D. and Read, P.A., Assessing future changes in extreme precipitation over Britain using regional climate model integrations. *International Journal of Climatology*, vol. 21, 2001, p. 1337-1356.

Lantuéjoul, C., *Geostatistical Simulation – Models and Algorithms*. Springer-Verlag, Berlin Heidelberg, 2002.

Prudhomme, C., and Reed, D.W., Mapping extreme rainfall in a mountainous region using geostatistical techniques: a case study in Scotland. *International Journal of Climatology*, vol. 19, 1999, p.1337-1356.

Thiessen, A.H., Precipitation averages for large areas. *Monthly Weather Review*, vol. 39, 1911, p. 1082-1084.

Wigley, T.M.L., Lough, J.M. Jones, P.D., Spatial patterns of precipitation in England and Wales and a revised, homogenous England and Wales precipitation series. *Journal of Climatology*, vol. 4, 1984, p. 1-25.

GEOSTATISTICAL INDICATORS OF WATERWAY QUALITY FOR NUTRIENTS

C BERNARD-MICHEL and C de FOUQUET
Ecole des Mines de Paris – Centre de Géostatistique
35, rue Saint Honoré 77305 Fontainebleau

Abstract. This paper aims at constructing geostatistical indicators to quantify water quality for nutrients. It presents a method to estimate the yearly mean and the 90th percentile of concentrations, taking into account temporal correlation, irregularity of sampling and seasonal variations of concentrations. On simulations, segment of influence declustering, kriging weighting and a linear interpolation of empirical quantiles are calculated and compared to the currently used statistical inference, based on the independence of random variables. These methods make it possible to correct the bias of the yearly mean and the quantile, and to improve their precision, giving a better prediction of the estimation variance. The study focuses on nitrates, in the Loire-Bretagne basin (France).

1 Introduction

In order to assess river water quality, nitrate concentrations are measured in different monitoring stations and summarized in a few synthetic quantitative indicators such as the 90% quantile of yearly concentrations or the annual mean, making it possible to compare water quality in different stations, and its yearly evolution. The current French recommendations are based on the water quality's evaluation system (SEQ EAU) and the water framework directive in Europe, which aims at achieving good water status for all waters by 2015. These calculations, however, use classical statistical inference, essentially based on a hypothesis proved to be incorrect for many parameters: time correlation is not taken into account. Moreover, the seasonal variations of concentrations and the monitoring strategy are ignored. Because of the streaming, nitrate concentrations are high in winter and low in summer (Payne, 1993), and then if sampling frequency is increased in time in winter out of precaution, the annual mean and the quantile are falsely increased. It is therefore necessary that the estimation takes into account both time correlations and irregularity of the measurements. We show that kriging or segment weights facilitate correction of the bias and improved assessment of the yearly temporal mean, as well as the quantile. For quantile estimation, the known bias of classical empirical calculations can be reduced using a linear interpolation of the empirical quantile function. Methods are presented and compared for simulations of nutrients.

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 907–912.

2 Annual concentrations at one station on the river Cher

Classical statistical inference consists of estimating the annual mean of nitrate concentrations by the sample mean and the 90% quantile by the empirical quantile. Figure 1 (left) and table 1 give an example of nitrate concentration measurements from the river Cher in 1985 in France. The annual mean and the 90th percentile have been estimated first with the totality of measurements (6 in summer, 12 in winter), then with an extracted sample of one regular measurement a month.

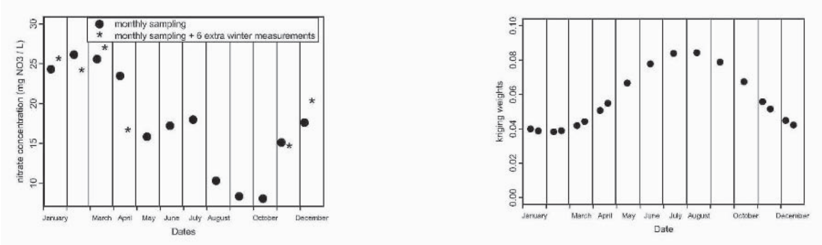


Figure 1. Preferential sampling of nitrates concentration during one year at one monitoring station (70300). Left: concentrations as a function of dates. Note that the frequency of measurements is doubled in winter. Right: associated kriging weights.

Table 1. Statistical yearly mean and quantile estimations corresponding to nitrates concentrations of Figure 2.

Sample size	Sample mean	90% quantile
12	17.49	25.56
18	18.75	26.12

Note that the sample mean is increased by 7% and quantile is increased by 3% when sampling is increased in winter. This difference, which can be much more important (up to 15 %) depending on the monitoring station, is a consequence of the temporal correlation (Figure 2, left). Because of this correlation, better methods are needed to assess the yearly temporal mean and the quantiles.

3 Methodology

3.1 THE YEARLY MEAN

For most of the monitoring stations, experimental temporal variograms calculated on nitrates concentrations show the evidence of a time correlation. For independent data, the sample mean (i.e., the arithmetic mean of the experimental data) is known to be an unbiased estimator; but in the presence of a time correlation, the sample mean is no longer unbiased, particularly when sampling is preferential. To correct this bias, two methods are studied:

- Kriging with an unknown mean (ordinary kriging, or OK), which takes into account correlation in the estimation of the annual mean and in the calculation of the estimation variance;

- A geometrical declustering, the only objective of which is to correct the irregularity of sampling.

These methods, detailed below, will be compared with simulations (Section 4):
a) In classical statistics (Saporta, 1990; Gaudoin, 2002): sample values $z_1, z_2, ..., z_n$ are interpreted as realizations of independent and identically distributed random variables $Z_1, Z_2, ...Z_n$ with expectation m. The yearly mean is estimated by the sample mean, denoted $m^* = \frac{1}{n}\sum_{i=1}^{n} Z_i$. The estimation variance $Var\left(m^* - m\right) = \sigma^2 / n$ is deduced from the experimental variance σ^2:

b) With temporal kriging (Matheron, 1970; Chilès and Delfiner, 1999): sample values are interpreted as a realization of a random correlated function $Z\left(t\right)$ at dates $t_1, t_2, ..., t_n$. In this situation the usual parameter of a distribution is not estimated, but rather the temporal, $Z_T = \frac{1}{T}\int_T Z\left(t\right)dt$, which is defined even in the absence of stationarity. Estimation proceeds using ordinary block kriging, as follows, $Z_T^* = \sum_{i=1}^{n} \lambda_i Z_i$ where λ_i are the kriging weights and the kriging variance is given by $Var(Z_T^* - Z_T)$
Analytical expressions necessary to calculate the kriging weights are easy to calculate in 1D, even without discretization. Figure 1 (right) gives an example of the kriging weights, assigning lower weights to winter values, which corrects the bias. The estimation variance and confidence interval, overestimated by classical statistics, are reduced by kriging taking into account the temporal correlation (figure 2, left) and the annual periodicity of the concentration.

c) Segment declustering, corresponding to 1D polygonal declustering (Chilès and Delfiner, 1999).

3.2 THE 90% QUANTILE OF NUTRIENTS CONCENTRATIONS

The 90% quantile is used to characterize high concentrations but the empirical quantile has proven to be biased even for independent realizations of the same random variable (Gaudoin, 2002). Moreover, it does not take into account temporal correlation and sample irregularity. In the literature there are many references pertaining to percentile estimation, especially with regard to statistical modelling of extreme values. However, many measurements are needed because of the appeal to asymptotic theorems. For an average of 12 measurements a year, classical non-parametric statistics will be used here, applying a linear interpolation of the empirical quantile. Then, to take temporal correlation into account, the data will be weighted using kriging or segment weights

4 Testing methods on simulations

4.1 CHOICE OF THE MONITORING STATION AND SIMULATION

Because complete time series for one year are not available, we use simulations for which the annual mean and the 90% quantile are known. Assuming that measurement per day corresponds to the daily concentration, we construct conditional simulations of "daily" concentrations with respect to the experimental data of real monitoring stations. Thus, we are able to compare the different estimations to the real yearly mean and quantile values. Different samplings schemes will be used: preferential, regular and irregular sampling.

The presentation is here restricted to station number 70300, on the river Cher, sampled monthly. One thousand of conditional simulations of 365 days conditioned by real measurements in 1985 were constructed (one simulation example is given on figure 2 (right)), using the fitted variogram presented on figure 2 (left). This experimental variogram calculated over several years reflects the annual periodicity of nitrate concentrations. Variograms calculated for each season would differ, but as we are interested in the global annual statistics, the averaged variogram on one year is sufficient (Matheron, 1970).

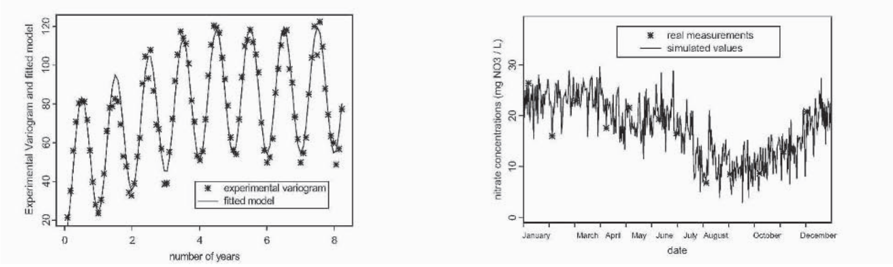


Figure 2. Monitoring station on the river Cher, nitrate concentration. Left: experimental variogram and fitted model, with a lag of 30 days. Right: one conditional simulation.

4.2 ESTIMATION OF THE YEARLY MEAN ON PREFERENTIAL SAMPLING

The results of the estimation of the yearly mean are presented In Figure 3 using 10 simulations. Estimations are compared on preferential sampling which comprises 18 measurements a year (6 values in summer and 12 values in winter) using monthly sampling (as in figure 1).

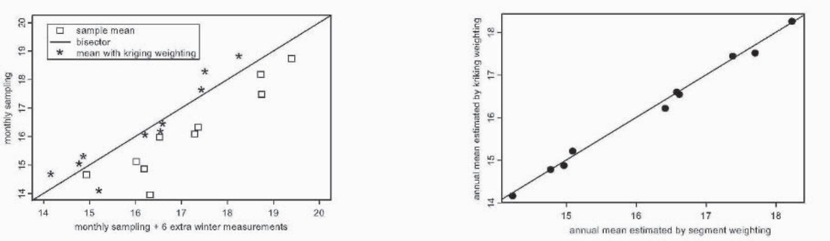


Figure 3. A comparison between the kriging and sample mean for the estimation of the yearly mean. On the left, scatter diagram between monthly sampling and preferential

sampling estimations by kriging and statistics. On the right, scatter diagram between the yearly mean estimated with kriging weights and segment weights.

Figure 3 (left) shows that kriging rectifies the bias of the sample mean in the case of preferential sampling. Moreover, the kriging variance is lower than the predicted statistical variance of the mean of independent variables, namely because of the yearly periodic component of the variogram. Here, the calculation of the estimation variance obtained with classical statistics is on average 70 % higher than that obtained with kriging. Figure 3 (right) shows the equivalence of segment and kriging weighting in the estimation of the yearly mean.

It can be concluded that kriging corrects the bias in case of preferential sampling, that it yields a better assessment of the yearly mean, and improved precision. However, if we are only interest in the value of the annual mean, segment declustering can be used because of its simplicity. If precision is needed, then kriging should be preferred.

4.3 QUANTILE ESTIMATION ON IRREGULAR SAMPLING

In this example, samples of different sizes have been extracted from each of the 1000 simulations, for 4 to 36 measurements a year, irregularly spaced in time. Thus we obtain 1000 samples of size 4, 1000 of size 5 etc…We estimate the 90% quantile for each sample using classical statistics and kriging or segment declustering on a linearly interpolated empirical quantile (Bernard-Michel and de Fouquet, 2004). Results are given in average for each different sample size and shown in Figure 4. They are compared to quantiles calculated in average on the 1000 simulations of 365 days.

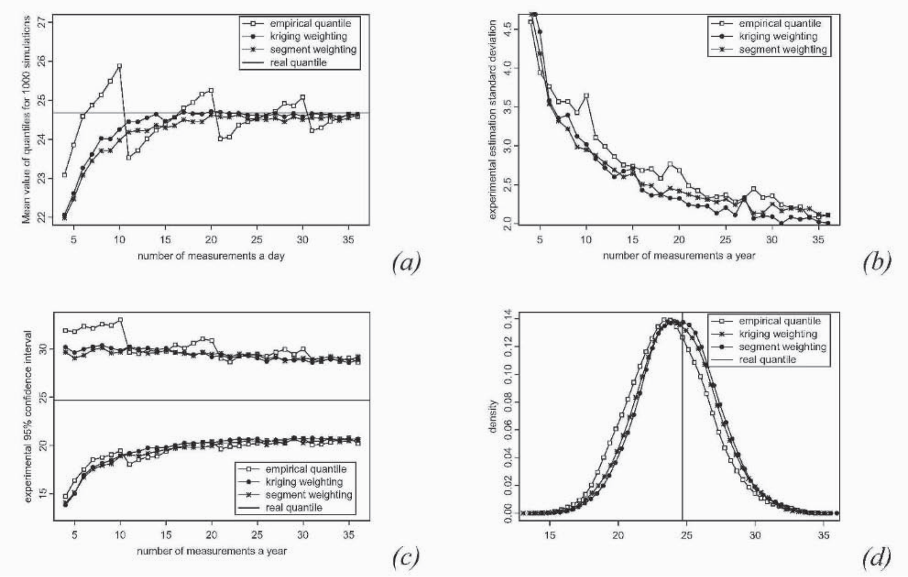


Figure 4. Average of the quantile estimation for temporal correlated concentrations, compared with the empirical quantity, for 1000 simulations. This empirical quantity corresponds to the mean, calculated on all the simulations, of the 90% quantiles of 365 values. All calculations are made by linear interpolation of quantiles. Upper left figure (a): average of quantiles estimation. Upper right figure (b): experimental estimation

standard error. Lower left figure (c): experimental 95% confidence interval. Lower right figure (d): histogram of quantile errors for samples of size 12.

Figure 4 (a) shows that the linear interpolation of the empirical quantile corrects the bias very well. Kriging or segment of influence weighting takes into account the sampling irregularity. However, the estimation variance (figure 8 (b)) remains larget. Actually, for 36 measurements a year, errors still represent approximately 8% of the real quantile which gives an approximate 95% confidence interval (figure 8. (c)) of ±20% around the real quantile because of the quasi normality distribution of errors (figure 8 (d)). For 12 measurements a year, they reach 11% of the real quantile, and 18% for 4 measurements a year. In the presence of time correlation, the theoretical estimator of a confidence interval would be difficult to construct. Even when random variables are independent, the theoretical interval (Gaudoin, 2002) is not satisfactory because it is limited by the higher order statistics. Simulations can be used to evaluate errors made in estimations and to determine the required sample size to achieve a desired precision.

5 Conclusions

Results are similar for other stations. Kriging the annual mean allows the bias induced by a preferential sampling of high concentration periods to be corrected. Associated with linear interpolation of the experimental quantile function, the kriging weights give an empirical estimation of quantiles that is practically unbiased. The segment of influence weighting can be used to simplify the calculations. In all cases, one or two measurements a month are not sufficient for a precise estimation of the yearly 90% quantile.

Acknowledgements

Thanks to the French Ministry of Environment (MEDD) for financing this study and L-C. Oudin and D. Maupas of the Agence de l'Eau Loire-Bretagne for their help.

References

Bernard-Michel, C. and de Fouquet, C., *Calculs statistiques et géostatistiques pour l'évaluation de la qualité de l'eau*. Rapport N-13/03/G. Ecole des Mines de Paris, Centre de géostatistique, 2003.

Bernard-Michel, C. and de Fouquet, C., *Construction of geostatistical indicators for the estimation of waterway's quality*, Geoenv, 2004.

Chilès J P., Delfiner P (1999) *Geostatistics. Modeling spatial uncertainty*. Wiley series in probability and statistics.

Gaudoin, O., *Statistiques non paramétriques*. ENSIMAG : notes de cours deuxième année, 2002 http://www-lmc.imag.fr/lmc-sms/Olivier.Gaudoin/

Littlejohn, C. and Nixon, S. and Cassazza, *Guidance on monitoring for the water framework directive*. Final draft. Water framework Directive, Working group 2.7, monitoring, 2002

Matheron, G., *La théorie des variables régionalisées, et ses applications*. Les cahiers du Centre de Morphologie Mathématique. Ecole des Mines de Paris. Centre de géostatistique, 1970.

Payne M. R. Farm (1993) Waste and nitrate pollution. Agriculture and the environment. John Gareth Jones. Ellis Horwood series in environmental management.

Saporta, G., Probabilités, analyse des données et statistiques, Technip, 1990.

APPLICATION OF GEOSTATISTICAL SIMULATION TO ENHANCE SATELLITE IMAGE PRODUCTS

CHRISTINE A. HLAVKA and JENNIFER L. DUNGAN
Ecosystem Science and Technology Branch, NASA Ames Research Center, Moffett Field, CA 94035-1000

Abstract. With the deployment of Earth Observing System (EOS) satellites that provide daily global imagery, there is increasing interest in defining the limitations of the data and derived products due to their coarse spatial resolution. Much of the detail, i.e. small fragments and notches in boundaries, is lost with coarse resolution imagery provided by systems such as the EOS MODerate Resolution Imaging Spectroradiometer (MODIS). Higher spatial resolution data such as the EOS Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER), Landsat Thematic Mapper and airborne sensor imagery provide more detailed information but are less frequently available.

There is, however, both theoretical and analytical evidence that burn scars and other fragmented types of land covers form self-similar or self-affine patterns, that is, patterns that look similar when viewed at widely differing spatial scales. Therefore small features of the patterns should be predictable, at least in a statistical sense, with knowledge about the large features. Recent developments in fractal modeling for characterizing the spatial distribution of undiscovered petroleum deposits are thus applicable to generating simulations of finer resolution satellite image products. We present example EOS products, an analysis to investigate self-similarity and a discussion of simulation approaches.

1 Terrestrial remote sensing with coarse spatial resolution

1.1 THE PROBLEM

The Earth Observing System (EOS) is a series of satellites launched by NASA during the past decade to make scientific measurements of the terrestrial land, oceans and atmosphere in order to understand how the Earth functions as a planet (http://eospso.gsfc.nasa.gov). Each satellite, with its suite of sensors collecting data in different parts of the electromagnetic spectrum, is meant to exploit aspects of surface and atmospheric interaction to make inferences about biophysical and geophysical variables. Maps of these variables are being used as model input for understanding how changes in the Earth's surface, such as deforestation and

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 913–919.

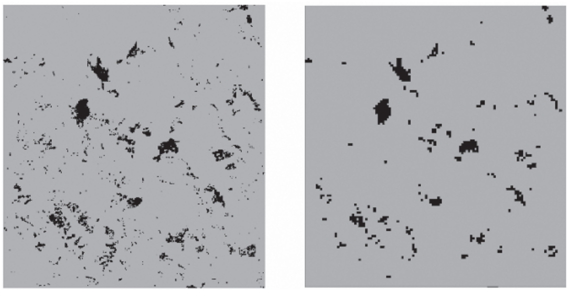


Figure 1. On the left, fire scars in the Brazilian savanna mapped with Landsat TM imagery. On the right, fire scars mapped with simulated MODIS imagery.

biomass burning are related to trends in atmospheric chemistry and climate. Other applications are being developed to address the priorities of other government agencies and sensor products are freely available to the general public.

In order to acquire global coverage at high temporal frequency, many EOS sensors acquire data at coarse spatial resolution (on the order of 1 km). This represents a one to three order of magnitude change in the size of the fundamental spatial unit of measurement (representing a huge difference in support) compared to other data sources such as Landsat Thematic Mapper (30 m), airborne sensors or ground observations (typically 1–15 m). Coarse products therefore lack details such as small fragments of a land cover type (Figure 1). These details are especially relevant to studies of land disturbances such as fire because land cover types such as burn scars and open water are often highly fragmented. Loss of detail due to coarse resolution effects limits the utility of these products and complicates quality assessment that typically involves comparison with finer resolution information. Most importantly, coarse resolution may lead to significant biases in estimating quantities of interest. For example, the areal extent of fragmented types of land cover might be underestimated due to omission of small fragments (Hlavka and Livingston, 1997; Hlavka and Dungan, 2002). Conversely, the proportion of more dominant land cover types might be overestimated because of missing small holes. There may be biases in non-linear computations in biogeophysical models because image values are pixel-average measures (in counts) of radiance averaged over the pixel, but $f(E(x))$ is not equal to $E(f(x))$ when f is non-linear (Dungan, 2001).

1.2 MOTIVATION

We address the coarse resolution problem by proposing novel methods for 1) adjusting area statistics derived from map products and 2) by considering how the products themselves might be "enhanced" to represent information from smaller supports (finer spatial resolutions) with better fidelity. Both approaches exploit the fractal nature of the phenomenon being mapped. This paper is a brief summary of our progress to date with both approaches. To develop the methods, we have used data from the MODerate Resolution Imaging Spectroradiometer (MODIS) sensor, in some cases with reference data from higher resolution Landsat images.

1.3 MODIS PRODUCTS

The MODIS instruments aboard NASA's Terra and Aqua satellites acquire image data with 250 m, 500 m, and 1000 m pixels in visible and infrared wavelengths. This imagery is therefore at a much coarser resolution than Landsat Thematic Mapper's (TM) 30 m pixels. The smaller data volume and broader swath width of MODIS allow for more frequent coverage and monitoring the Earth's surface in a more comprehensive manner, both spatially and temporally, as required for the scientific understanding of global climate issues. Unlike Landsat TM or other older types of satellite imagery, the MODIS imagery is processed into high-level products such as digital maps of land cover type and percent forest cover (Justice et al., 2002) representing both categorical and continuous variables. "Land cover" (Friedl et al., 2002) and "burn area" (Roy et al., 2002) products are both categorical. The land cover product is a map of land cover type (forest, shrubland, bare ground, wetlands, etc.) with 1 km pixels. The burned area product maps burn scars by detection date with 500 m pixels. The "vegetation continuous fields" product maps percent coverage of forest and non-forest vegetation types with 500 m pixels (Hansen et al., 2002). Among the data sets we have worked with are land cover and percent forest over an area of boreal forest and lakes in Manitoba and the burned area product in the Okavango Swamp in Botswana, Africa where there are massive seasonal fires.

There is fundamental aspect of satellite imagery as data that is central to its interpretation and processing. Most data collected on the ground and analyzed with geostatistical methods are essentially point data, that is, the support is very small compared to distances between samples. Satellite imagery, on the other hand, is synoptic. The pixel sizes are approximately the distance between neighboring pixel centers, so that the pixels tile the scene being observed from the satellite platform. Pixel values represent an average over the extent of the pixel. In image-based products, values represent a transformation of pixels values (e.g. an estimate of percent forest) or are a code for a predominant category, such as land of a type (e.g. forest) or condition (e.g. recently burned).

2 Fractal properties of geographic features observable from space

Analysis of geographic data has provided evidence of the fractal nature of the Earth's surface and a variety of geographic features. Probably the best-known example is the coastline of Britain whose length is a power function of the length of the measuring stick (Mandelbrot, 1982). The sizes of islands, lakes, and patches of burned vegetation have been shown to have fractal (i.e. Pareto or power) distributions (Mandelbrot, 1982; Korcák, 1938; Malamud et al., 1998) with probability density function (theoretical normalized histogram) of the form:

$$p(x) = Prob(x < X)/dx = Ax^{-B} \; for \; x > \epsilon > 0, \; A > 0, \; B > 0 \qquad (1)$$

where ϵ is a value close to zero. Fractal patterns in continuous field data such as percent forest cover might also be expected because such measures are strongly related to burning – forest age in many regions is typically determined by when it last burned.

The fractal nature of landscape features has also been observed with satellite imagery. For example, Ugandan forest boundaries are approximately a power function of satellite pixel size (Hlavka and Strong, 1992). Size distributions of burn scars and water bodies have been found to be approximately fractally distributed (Hlavka and Dungan, 2002). However, sometimes the distribution of feature sizes are more lognormal than fractal, possibly due to imaging effects besides non-detection of fragments below the pixel size.

Patterns of small and large fragments are linked through process models, in particular those associated with self-organized criticality (Malamud et al., 1998; Hergarten, 2002) and other landscape models that generate fractal patterns of disturbances. The variogram $\gamma(h)$ of a self-similar fractal pattern $Z(x)$ is:

$$\gamma(h) = E[Z(x+h) - Z(x)]^2 \sim h^{2H} \tag{2}$$

where H is known as the Hausdorff exponent (Hergarten, 2002). The Fourier spectrum follows a similar form with amplitudes $P(\nu) \sim |\nu|^{\beta}$ where β is the spectral exponent (Hergarten, 2002; Chen et al., 2001). These characterizations of fractal pattern at a wide range of spatial resolutions are the basis for proposing adjusted area estimates and "enhanced" versions of satellite image products that have the original large land cover patterns from the coarse resolution product with simulated smaller details.

3 Solutions to the coarse resolution problem using fractal properties

3.1 ADJUSTING AREA STATISTICS

Models of size distribution have been experimentally used to adjust area estimates for missing small burn scars and water bodies following Maxim and Harrington (1982) with simulated and actual MODIS products and with Advanced Very High Resolution Radiometer (AVHRR) weather satellite imagery (1 km pixels). Adjusted burn area estimates based on simulated MODIS or AVHRR were found to be closer to Landsat estimates than simple count × pixel area (Hlavka and Dungan, 2002). A similar approach has been used to address petroleum reserve estimation (Barton and Scholz, 1995).

Quantile-quantile plots of the observed sizes of burn scars in the Okavango burned area product and patches of wetland and open water in the Manitoba land cover product indicated a fit to a fractal or lognormal distribution, as indicated by the degree to which the trend in the data fit a straight line (Figure 2). Area adjustment was implemented with a program "ltfill" (http://lib.stat.cmu.edu/s) that uses truncated data ($x > C$, a user- specified value below which data are missing or unreliable) to estimate the parameters of a lognormal, fractal, or exponential distribution, then estimates the number and sum of all values. For this

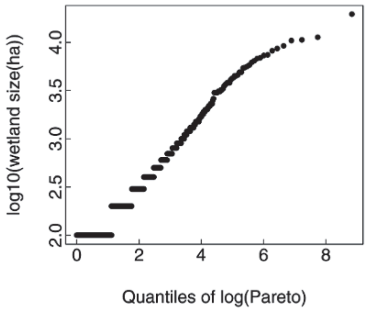


Figure 2. Quantile-quantile plot comparing the observed quantiles of sizes of boreal wetlands mapped on the MODIS land cover product to theoretical quantiles of the fractal distribution.

application, the cutoff value C is chosen to be the area of one or two pixels, since fragments smaller than this cannot be reliably detected.

For boreal wetlands, the adjusted area was half again the usual area estimate. The Okavango burns were extensive with many fragments of unburned area within the largest burn scars, so burned area was adjusted for both small, unobserved burns and unobserved, unburned fragments with the result of slightly decreasing the areal estimate.

3.2 SIMULATING 2D FIELDS AT FINE SPATIAL RESOLUTION

Beyond simple adjustments to global statistics, it is desirable to have a model of what the actual 2D field of the variable of interest looks like. This raises a typical geostatistical problem but with a new twist: how to create synthetic realizations defined on a small support that have the correct statistical properties of the phenomenon of interest without measurements on small support. For lack of a better term, we use the word, “enhancement” to refer to the process of modeling a 2D field with small support given a 2D field with large support. Simulation using such a model would be used in an unconditional sense in that there are no measurements to honor at the small support. The key information needed to create an objective function or other figure of merit for constraining the simulations is the extrapolation of the spatial covariance to smaller lags using (2). The simulation must also incorporate elements that play the role of conditional information so that the locations of known features are correct.

One approach is the use of simulated annealing (Deutsch and Cockerham, 1996). The concept would be to use the coarse resolution image initially, then “anneal” it using an objective function defined at least partly by the inferred spatial covariance properties of a fine resolution image based on the fractal characteristics of the coarse resolution image. Another approach might be an adaptation of a cascade model (Cheng, this volume), wherein new pixel values replacing an original pixel value V are constrained to have means equal to V and variances consistent with extrapolated short distance covariances or constraints based on multiple point statistics (Arpat and Caers, this volume).

The approach we took was to generate unconditional realizations in the frequency domain using a modified version of spectral simulation with phase identification (Yao, 1998; Yao et al., this volume). This method takes advantage of the deterministic relationship between spatial covariance $C(h)$ and the magnitude of the Fourier transform $A(j) : FT(C(h)) = |A(j)|^2$ and the role of the Fourier transform phase in determining the location of features in an image. This approach was tested using percent forest layer of the vegetation continuous fields product. We selected a 128×128 pixel chip from this product for an area in central Canada. The chip was of an area near Lake Athabasca and showed a blocky pattern related to forest clear-cuts of various ages and stages of regrowth. The magnitudes of the Fourier transform ($|FT|$) along three lines (the x and y axes and the 45 degree line) showed power curve trends, with similar exponents, consistent with an isotropic fractal pattern. The $|FT|$ was quite rough, i.e. deviations from the trend around $log|FT|$ versus $log(frequency)$ were large. We created a 1024 × 1024 pixel version of the original chip by taking the FT inverse of a 1024 × 1024 FT created by extrapolating the original $|FT|$ into higher frequencies according to a regression power model and padding out the original phases with zeroes in higher frequencies. The resulting 1024 × 1024 data product had high amplitude noise and periodic features (probably an artifact of processing) that obscured the pattern in the original 128 × 128 chip, indicating that changes in procedure need to be made before a realistic simulation is achieved.

4 Conclusions

We have presented work to address the biases that may arise from the use of coarse resolution satellite data for mapping fragmented landscape phenomena such as burn scars, water bodies and forest patches. Adjusting global area estimates using a fractally-based extrapolation of the size distribution is a promising technique that requires further testing on a large set of representative data. We have not yet succeeded in creating realistic enhanced images using a simulation approach. Geostatistical methods of unconditional simulation should be further tested for this purpose. Results from such an effort have the potential to reduce discrepancies in global biogeophysical measurements for many important Earth-system science issues.

Acknowledgements

Support for this work was provided by NASA's Earth Science Enterprise Office of Research. Burned area products supplied by David Roy, University of Maryland. We also thank Rajkumar Thirumalanambi for his participation.

This page intentionally blank

References

Arpat, B.G. and J Caers, A Multiple-scale, pattern-based approach to sequential simulation, this volume.

Barton, C.C. and C. H. Scholz, The fractal size and spatial distribution of hydrocarbon accumulations: Implications for resource assessment and exploration strategy, 1995. In: C.C. Barton and P.R. LaPointe, Eds., *Fractals in Petroleum Geology and Earth Processes*, pp. 13–34, Plenum Press, 1995.

Chen Z., Osadadetz K. and Hannigan P., An improved model for characterizing spatial distribution of undiscovered petroleum accumulations, in *Proceedings of the Annual Conference of the International Association for Mathematical Geology*, Cancun Mexico, Sept. 6-12, 2001.

Cheng, Q. A New model for incorporating spatial association and singularity, this volume.

Deutsch, C. and Cockerham, P., The application of simulated annealing to stochastic reservoir simulation, Mathematical Geology, 26, 67–82, 1996.

Dungan J.L., Scaling up and scaling down: relevance of the support effect on remote sensing of vegetation, in Tate and Atkinson (editors), *Modelling Scale in Geographical Information Science*, pp. 231–235, John Wiley and Sons, 2001.

Friedl M.A., McIver D.K., Hodges J.C.F., Zhang X.Y., Muchoney D., Strahler A.H., Woodcock C.E., Gopal S., Schneider A., Cooper A., Baccini A., Gao F. and Schaaf C., Global land cover mapping from MODIS: algorithms and early results, *Remote Sensing of Environment*, 83, 287–302, 2002.

Hansen, M.C., DeFries, R.S., Townshend, J.R.G., Sohlberg, R., Dimiceli, C. and Carroll M., Towards an operational MODIS continuous field of percent tree cover algorithm: examples using AVHRR and MODIS data, *Remote Sensing of Environment*, 83, 303–319, 2002.

Hergarten S., *Self-organized criticality in earth systems*, Springer-Verlag, Berlin, 2002.

Hlavka C.A. and Dungan J.L., Areal estimates of fragmented land cover – effects of pixel size and model-based corrections, *International Journal of Remote Sensing*, 3, 711–724, 2002.

Hlavka C.A. and Livingston J., Statistical models of landscape pattern and the effects of coarse resolution satellite imagery on estimation of area, *International Journal of Remote Sensing*, 18, 2253–2259, 1997.

Hlavka C.A. and Strong L., Assessing deforestation and habitat fragmentation in Uganda using satellite observations and fractal analysis, *Journal of Imaging Science and Technology*, 36, 440–445, 1992.

Justice C.O., Townshend J.R.G., Vermote E.F., Masuoka E., Wolfe R.E., Saleous N.E., Roy D.P. and Morisette J.T., An overview of MODIS land data processing and product status, *Remote Sensing of Environment*, 83, 3–15, 2002.

Korcák J., Deux types fondementaux de distribution statistique, *Bulletin de L'Institute International de Statistique*, III, 295–299, 1938.

Malamud B.D., Morein G. and Turcotte D.L., Forest fires: an example of self-organized critical behavior, *Science*, 281, 1840–1842, 1998.

Mandelbrot B.B., *The Fractal Geometry of Nature*, Freeman, San Francisco, 1982.

Maxim L.D. and Harrington L., Scale-up estimators for aerial surveys with size dependent detection, *Photogrammetric Engineering and Remote Sensing*, 48, 1271–1287, 1982.

Roy D., Lewis P. and Justice C., Burned area mapping using multi-temporal moderate spatial resolution data - a bi-directional reflectance model-based expectation approach, *Remote Sensing of Environment*, 83, 263–286, 2002.

Yao, T., Conditional spectral simulation with phase identification, *Mathematical Geology*, 30, 285–108, 1998.

Yao, T., C. Calvert, G. Bishop, T. Jones, Y. Ma and L. Foreman, Spectral component geologic modeling: A new technology for integrating seismic information at the correct scale, this volume.

GEOSTATISTICAL NOISE FILTERING OF GEOPHYSICAL IMAGES: APPLICATION TO UNEXPLODED ORDNANCE (UXO) SITES

HIROTAKA SAITO[1], TIMOTHY C. COBURN[2] and SEAN A. MCKENNA[1]
[1]*Geohydrology Department, Sandia National Laboratories, P.O. Box 5800, MS 0735, Albuquerque, NM, 87185-0735*
[2]*Department of Management Science, Abilene Christian University, ACU Box 29315, Abilene, TX, 79699*

Abstract. Geostatistical and non-geostatistical noise filtering methodologies, factorial kriging and a low-pass filter, and a region growing method are applied to analytic signal magnetometer images at two UXO contaminated sites to delineate UXO target areas. Overall delineation performance is improved by removing background noise. Factorial kriging slightly outperforms the low-pass filter but there is no distinct difference between them in terms of finding anomalies of interest.

1 Introduction

The goal of unexploded ordnance (UXO) site characterization is to delineate target areas, within which UXOs are expected to be clustered and required excavation or further investigation from spatially exhaustive geophysical information. The most straightforward characterization approach consists of applying a threshold to each pixel in the geophysical map; values above the threshold are considered as potential UXO or objects of interest. However, since available geophysical maps are not usually smooth, this approach can lead to a noisy mosaic of pixels flagged for excavation. This research aims to improve the final decision map (i.e., binary map) by filtering spatially uncorrelated background noise from the geophysical map using factorial kriging (Wen and Sinding-Larsen, 1997). A region growing method is then applied to the smoothed map to detect UXO target areas. The basic idea of the region growing method is to start from a point that meets a criterion (e.g., highest geophysical signal value) and to extend the area by adding adjacent pixels in all directions until a specified number of pixels are included or a boundary is detected (Hojjatoleslami and Kittler, 1998). It can then provide an optimal estimate of the target area, which is of great regulatory interest. The influence on the final decision maps due to the search window size in factorial kriging is investigated and results are compared to those obtained with a non-geostatistical filtering technique (i.e., low-pass filter).

In this study, two UXO sites are used: a hypothetical site created with a Poisson simulator (McKenna, et al., 2001) and the Pueblo of Isleta site in New Mexico, where an exhaustive magnetic analytic signal map is available. The benefit of using the hypothetical site is that the true spatial distribution of objects is known so that any type

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 921–926.

of investigation is possible and the accuracy and precision of the results can be fully evaluated. At the hypothetical site, the spatial distribution of objects is modelled by a non-homogeneous Poisson process (McKenna, et al., 2001) and corresponding analytic signal values are also simulated. Background noise, on the other hand, was modelled by a homogeneous Poisson process (i.e., random noise), in which the object density is uniform across the site. The size of the hypothetical site is 5000 × 5000 meters and the single UXO target is located in the center of the site (Figure 1 left). A 50 × 50 meter pixel is used as the spatial support over which any characterization decision is made. There are a number of simulated objects in each pixel but only the largest signal value within each pixel is retained as a representative value of the pixel for further investigation (Figure 1, center).

The Pueblo of Isleta site has been surveyed by the Geophysical team of Oak Ridge National Laboratory using their airborne UXO detection system (Doll, et al., 2003). In their system, which is referred to as the Oak Ridge Airborne Geophysical System – Hammerhead Array (ORAGS^{TM}-HA), 8 magnetometers are deployed inside booms mounted to a helicopter. Figure 1 (right) shows the highest analytic signal value in 15 × 15 m pixels at the site. Since no excavation has been conducted at the site, a complete analysis is not possible.

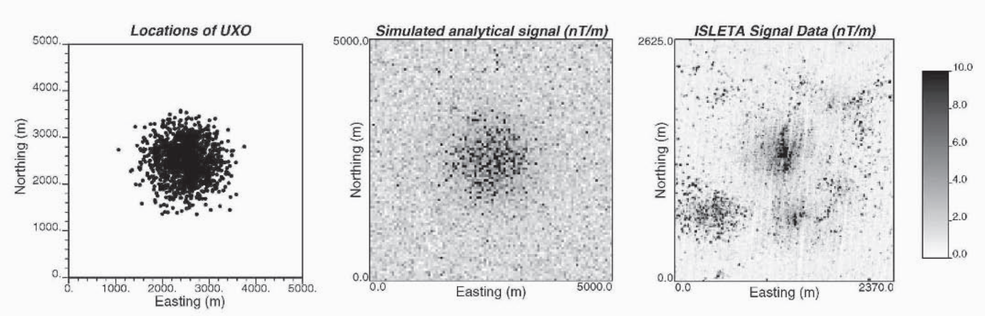


Figure 1. Simulated analytic signals [nT/m] at the hypothetical site (middle) with the true UXO distribution (left). The right image shows the analytic signal values (right) obtained at the Pueblo of Isleta site (S3) in New Mexico by Oakridge National Laboratory (2002).

2 Methods

This section briefly reviews the geostatistical filtering technique, factorial kriging, and the region growing method. For the non-geostatistical filtering approach, a low-pass filter is used with two different window sizes: 3 × 3 and 5 × 5.

2.1 FACTORIAL KRIGING

Factorial kriging (FK) is an algorithm to decompose an attribute into spatial components with different spatial scales, say (L+1) different scales (Goovaerts, 1997). The underlying assumption is that spatial components from different sources (e.g. scales) are independent of each other and they are additive. Under these conditions, the random function $Z(\mathbf{u})$ can be written as a sum of $(L+1)$ independent random functions and a single mean or trend $m(\mathbf{u})$:

$$Z(\mathbf{u}) = \sum_{l=0}^{L} Z_l(\mathbf{u}) + m(\mathbf{u})$$

Under this condition, the semivariogram of the random function $Z(\mathbf{u})$ is modeled as the sum of $(L+1)$ semivariograms of random functions $Z_l(\mathbf{u})$ (i.e., spatial components):

$$\gamma(\mathbf{h}) = \sum_{l=0}^{L} \gamma_l(\mathbf{h}) = \sum_{l=0}^{L} b_l\, g_l(\mathbf{h}) \qquad \text{with } b_l \geq 0$$

where b_l is the variance and $g_l(\mathbf{h})$ is the basic semivariogram model of the corresponding l-th random function $Z_l(\mathbf{u})$ with a certain spatial scale. By convention, the superscript l=0 denotes the nugget component, which is a spatially uncorrelated random function and is usually regarded as noise. Spatial components at different scales can be then easily identified from their experimental semivariograms.

When FK is applied to exhaustive data (e.g., images), it amounts to decomposing an observation or a map into individual components with different spatial scales, because of the exactitude property of the kriging estimator (Goovaerts, 1997, pp. 165). It allows one to filter out an l_0-th component (e.g., the nugget effect) from the observation and directly estimate the filtered value as a linear combination of surrounding data using:

$$w_{l_0}^{*}(\mathbf{u}) = \sum_{\alpha=1}^{n(\mathbf{u})} \nu_{\alpha l_0}(\mathbf{u})\, z(\mathbf{u}_\alpha)$$

The kriging weights, $\nu_{\alpha l0}(\mathbf{u})$, are obtained by solving the following system:

$$\begin{cases} \displaystyle\sum_{\beta=1}^{n(\mathbf{u})} \nu_{\beta l_0}(\mathbf{u})\, \gamma(\mathbf{u}_\alpha - \mathbf{u}_\beta) + \mu_{l_0}(\mathbf{u}) = \gamma(\mathbf{u}_\alpha - \mathbf{u}) - \gamma_{l_0}(\mathbf{u}_\alpha - \mathbf{u}) & \alpha = 1,\ldots,n(\mathbf{u}) \\ \displaystyle\sum_{\beta=1}^{n(\mathbf{u})} \nu_{\beta l_0}(\mathbf{u}) = 1 \end{cases}$$

The only difference from the ordinary kriging system is that the right-hand-side semivariogram is computed by subtracting the semivariogram of the l_0-th spatial component.

2.2 REGION GROWING METHOD

There are several algorithms available for the region growing method. The basic idea is to start from a point that satisfies a specific criterion (e.g. highest signal value) and to extend the area by adding adjacent pixels in all directions until a boundary is detected or a stopping rule satisfied. This method is based on the idea that pixels belonging to the same region have similar values or at least have values different from those in the background. Variants of the region growing method have different growing and stopping rules. In this study, the method proposed by Hojjatoleslami and Kittler (1998) is modified and used because of its theoretical simplicity and the robustness of the algorithm. The method proceeds as follows:

1. An arbitrary pixel within a region of interest or a pixel that satisfies some criterion (such as the pixel with the highest signal value) is selected. A spatially contiguous set of selected pixels is referred to as a *region*.
2. The pixel with the highest value among pixels adjacent to the *region* is added to the *region*.
3. Steps 1 and 2 are repeated until a given number of pixels are added to the *region*.

For demonstration purpose, the extension of the *region* is stopped when a specified number of pixels are included.

3 Results

Factorial kriging (FK) requires a linear model of regionalization. Figure 2 (left) shows an experimental semivariogram of analytic signal values of the hypothetical site and a fitted Gaussian model. The nugget component, which is usually related to spatial variability in very short scales, is then filtered out using FK (Figure 2, right). As expected, the image is smoothed out compared to the original signal image (Figure 1) especially in the background.

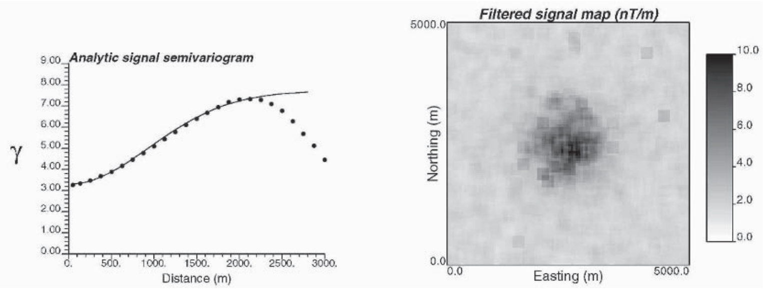


Figure 2. Experimental semivariogram of analytic signal values at the hypothetical site with a Gaussian model fit. Factorial kriging (FK) is used to remove the nugget component from the original analytic signal image. 25 neighbours are used in FK.

The region growing method is then applied to the filtered image to delineate a UXO target region. Figure 3 shows the *region* consisting of 1000 pixels obtained using both original (left) and filtered (right) images. Since the filtered image is much smoother than the original image, the boundary of the *region* for the filtered image is much more continuous. Considering the effort required for UXO excavation, the *region* selected using the original image is too noisy and not practical.

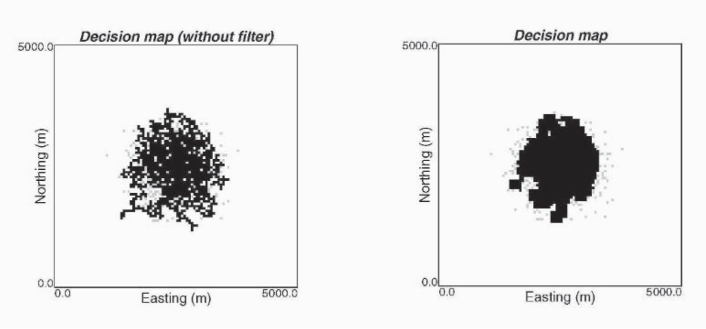


Figure 3. *Regions* (black pixels) selected using the original and filtered images by applying the growing region method. The extension of the region is stopped when 1000 pixels are included in the region. Gray pixels are not included in the *region* but contain at least one UXO (i.e., false negative).

The performance of the noise filtering technique and the region growing method is quantified by calculating the proportion of UXO, or anomalies of interest, that fall in the selected *region*. Table 1 summarizes the results for both sites. If the *region* with 1000 pixels is used at the hypothetical site, filtering the image (i.e., removing noise) does not improve identification of UXO. On the other hand, when 2000 pixels are included in the *region*, noise filtering slightly improves the delineation performance, especially when a larger search window is used.

Filtering	Hypothetical site		Isleta site	
	1000 pixels	2000 pixels	> 3 [nT/m]	> 10 [nT/m]
None	84.49	97.25	71.45	71.95
FK (*n*(**u**) = 25)	84.89	98.76	76.98	86.40
FK (*n*(**u**) = 9)	84.89	98.49	77.78	84.99
Low-pass (5×5)	84.89	98.76	71.74	83.00
Low-pass (3×3)	84.89	98.49	77.60	85.84

Table 1. Percentage (%) of UXO at the hypothetical site or anomalies of interest at the Isleta site located within the selected *region*. The underline values indicate the best results among different filtering techniques.

When the hypothetical site is used, only one region is considered. As for the Isleta site, since there seems to be more than one potential UXO target areas (Figure 1, right), the application of the region growing method is slightly different. To delineate more than one target area, the region growing method needs to be applied several times. For the Isleta site, since there are five suspected areas, five starting pixels are chosen from each suspected area. Each region is then grown to a maximum of 1500 pixels in this example. When the *region* selects a pixel already selected by a different *region* while growing, it stops. In addition, since there is no information about an actual UXO distribution at the Isleta site, two signal threshold values, 3.0 and 10.0 [nT/m], are considered to define "anomalies of interest." For both cases, the best results are obtained when FK is used to filter the background noise from the image. In general, the delineation of anomalies

This page intentionally blank

with higher signal values improves for both geostatistical and non-geostatistical filtering techniques compared to the result without any filtering (Table 1).

4 Conclusions

This study demonstrates a procedure to improve the delineation of suspected UXO target areas from exhaustive geophysical images. Two UXO contaminated sites, the hypothetical site and the Pueblo of Isleta site, are used. The basic approach is first to remove the background noise by smoothing the image. Then, the region growing method is applied to delineate the area with high signal values, which should be related well to UXO. This area is considered a suspected UXO target area. In this study, two different filtering techniques, factorial kriging (geostatistical technique) and low-pass filter (non-geostatistical technique), are compared.

The performance of the approach is quantified by calculating the proportion of UXO or anomalies of interest found in the *region* selected. Both filtering techniques improve the overall delineation performance. Factorial kriging is superior at the Isleta site, while there is no distinct difference between FK and low-pass filter at the hypothetical site in terms of finding UXO. One of the main reasons is that, as the range of semivariogram (> 2000m) is much larger than the filtering window size (250m); factorial kriging is almost identical to takeing the local average as done in the low-pass filtering procedure. The geostatistical approach is expected to perform better when the filtering window size is much closer to the range of the semivariogram.

Acknowledgements

This study was financially supported by the Strategic Environmental Research and Development Program (SERDP) UXO Cleanup program under grant UX-1200. The authors appreciate William E. Doll and Jeffery Gamey of Oak Ridge National Laboratory for providing the analytic signal data at the Pueblo of Isleta site. Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

References

Doll, W.E., Gamey, T.J., Beard, L.P., Bell, D.T., and Holladay, J.S., Recent advances in airborne survey technology yield performance approaching ground-based surveys, *The Leading Edge*, May, 2003, 420-425.

Goovaerts, P., *Geostatistics for Natural Resources Evaluation*, Oxford University Press, 1997.

Hojjatoleslami, S.A., and Kittler, J., Region growing: A new approach, *IEEE Transactions on Image Processing*, vol. 7, no. 7, 1998, p. 1079-1084.

McKenna, S.A., Saito, H., and Goovaerts, P., 2001, Bayesian approach to UXO site characterization with incorporation of geophysical information, *SERDP Project UX-1200*, Deliverable.

Wen, R., and Sinding-Larsen, R., Image filtering by factorial kriging - sensitivity analysis and application to Gloria side-scan sonar images, *Mathematical Geology*, vol. 29, no. 4, 1997, p. 433-468.

THEORY & SELECTED TOPICS

MODELING SKEWNESS IN SPATIAL DATA ANALYSIS WITHOUT DATA TRANSFORMATION

PHILIPPE NAVEAU$^{(1,2)}$ and DENIS ALLARD$^{(3)}$
(1) Dept. of Applied Mathematics, University of Colorado, Boulder, USA
(2) Laboratoire des Sciences du Climat et de l'Environnement, IPSL-CNRS, Gif-sur-Yvette, France
(3) INRA, Unité de Biométrie, Site Agroparc, 84914 Avignon, France

Abstract. Skewness is present in a large variety of spatial data sets (rainfalls, winds, etc) but integrating such a skewness still remains a challenge. Classically, the original variables are transformed into a Gaussian vector. Besides the problem of choosing the adequate transform, there are a few difficulties associated with this method. As an alternative, we propose a different way to introduce skewness. The skewness comes from the extension of the multivariate normal distribution to the multivariate skew-normal distribution. This strategy has many advantages. The spatial structure is still captured by the variogram and the classical empirical variogram has a known moment generating function. To illustrate the applicability of such this new approach, we present a variety of simulations.

1 Introduction

The overwhelming assumption of normality in the multivariate Geostatistics literature can be understood for many reasons. A major one is that the multivariate normal distribution is completely characterized by its first two moments. In addition, the stability of multivariate normal distribution under summation and conditioning offers tractability and simplicity. However, this assumption is not satisfied for a large number of applications. In this work, we propose a novel way of modeling skewness for spatial data by working with a larger class of distributions than the normal distribution. This class is called *general multivariate skew-normal distributions*. Besides introducing skewness to the normal distribution, it has the advantages of being closed under marginalization and conditioning. This class has been introduced by Domínguez-Molina et al., 2003 and is an extension of the multivariate skew-normal distribution first proposed by Azzalini and his coworkers (Azzalini, 1985, Azzalini, 1986, Azzalini and Dalla Valle, 1996 and Azzalini and Capitanio, 1999). These distributions are particular types of generalized skew-elliptical distributions recently introduced by Genton and N. Loperfido, 2005, i.e. they are defined as the product of a multivariate elliptical density with a

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 929–937.

skewing function. This paper is organized as follows. In Section 2, the definition of skew-normal distribution is recalled and notations are introduced. In Section 3, we first recall the basic framework of spatial statistics and then present the spatial skewed Gaussian processes. The estimation procedure of the variogram and skewness parameters is presented in details and illustrated on simulations. We conclude in Section 4.

2 Skew-normal distributions

Multivariate skew normal distributions are based on the normal distribution but a skewness is added to extend the applicability of the normal distribution while trying to keep most of the interesting properties of the Gaussian distribution. Today, there exists a large variety of skew-normal distributions (Genton, 2004) and they have been applied to a variety of situations. For example, Naveau et al., 2004, developped a skewed Kalman filter based on these distributions. In a Gaussian framework, spatial data are analyzed using the skew normal distribution (Kim and Mallick, 2002), but without a precise definition of skew normal spatial processes. As it will be shown in Section 3.1, this model leads to a very small amount of skewness and therefore is not very usefull in practice.

From a theoretical point of view, we use in this work the multivariate closed skew-normal distribution (Domínguez-Molina et al., 2003, González-Farías et al., 2004). It stems from the ”classical” skew-normal distribution introduced by Azzalini and its co-authors. It has the advantages of being more general and having more properties similar to the normal distribution than any other skew-normal distributions. A drawback is that notations can become cumbersome. The book edited by Genton, 2004, provides an overview of the most recent theoretical and applied developments related to the skewed distributions.

An n-dimensional random vector $\mathbf{Y}$ is said to have a multivariate closed skew-normal distribution denoted by $\mathrm{CSN}_{n,m}(\mu, \mathbf{\Sigma}, \mathbf{D}, \nu, \mathbf{\Delta})$, if it has a density function of the form:

$$c_m\ \phi_n(\mathbf{y}; \mu, \mathbf{\Sigma})\ \Phi_m(\mathbf{D}^t(\mathbf{y}-\mu); \nu, \mathbf{\Delta}),\ \text{ with } c_m^{-1} = \Phi_m(0; \nu, \mathbf{\Delta} + \mathbf{D}^t\mathbf{\Sigma}\mathbf{D}), \tag{1}$$

where $\mu \in \mathbb{R}^n$, $\nu \in \mathbb{R}^m$, $\mathbf{\Sigma} \in \mathbb{R}^{n\times n}$ and $\mathbf{\Delta} \in \mathbb{R}^{n\times 1}$ are both covariance matrices, $\mathbf{D} \in \mathbb{R}^{m\times n}$, $\phi_n(\mathbf{y}; \mu, \mathbf{\Sigma})$ and $\Phi_n(\mathbf{y}; \mu, \mathbf{\Sigma})$ are the n-dimensional normal pdf and cdf with mean μ and covariance matrix $\mathbf{\Sigma}$, and $\mathbf{D}^t$ is the transpose matrix of $\mathbf{D}$. When $\mathbf{D} = 0$, the density (1) reduces to the multivariate normal one, whereas it is equal to Azzalini's density (Azzalini and Dalla Valle, 1996), i.e. the variable $\mathbf{Y}$ follows a $\mathrm{CSN}_{n,1}(\mu, \mathbf{\Sigma}, \alpha, 0, 1)$, where α is a vector of length n. This distribution was the first multivariate skew-normal distribution and it was introduced by Azzalini and his coworkers (Azzalini, 1985, Azzalini, 1986, Azzalini and Dalla Valle, 1996 and Azzalini and Capitanio, 1999).

The $\mathrm{CSN}_{n,m}(\mu, \mathbf{\Sigma}, \mathbf{D}, \nu, \mathbf{\Delta})$ distribution defined by (1) is generated from the following bivariate vector. Let $\mathbf{U}$ be a Gaussian vector of dimension m and let us consider the augmented Gaussian vector $(\mathbf{U}^t, \mathbf{Z}^t)^t$ with the following distribution:

$$\begin{pmatrix} \mathbf{U} \\ \mathbf{Z} \end{pmatrix} \overset{\mathrm{d}}{=} N_{m,n}\left(\begin{pmatrix} \nu \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \mathbf{\Delta} + \mathbf{D}^t\mathbf{\Sigma}\mathbf{D} & -\mathbf{D}^t\mathbf{\Sigma} \\ -\mathbf{\Sigma}\mathbf{D} & \mathbf{\Sigma} \end{pmatrix} \right) \tag{2}$$

where $\overset{\mathrm{d}}{=}$ corresponds to the equality in distribution. Then, it is straightforward to show that conditional on $\mathbf{U} \leq \mathbf{0}$ the random vector $\mu + [\mathbf{Z}|\mathbf{U} \leq \mathbf{0}]$ is distributed according to a $\mathrm{CSN}_{n,m}(\mu, \mathbf{\Sigma}, \mathbf{D}, \nu, \mathbf{\Delta})$ as defined in Equation (1). Here the notation $\mathbf{A} \leq \mathbf{B}$ corresponds to $A_i \leq B_i$, for all $i = 1, \ldots, n$. This construction offers a wide range of possible models depending on the choice of μ, ν, $\mathbf{\Delta}$, $\mathbf{\Sigma}$ and $\mathbf{D}$. For more details on this type of construction, we refer to (Domínguez-Molina et al., 2003).

It is well known that the conditional vector $[\mathbf{Z}|\mathbf{U}]$ is also a Gaussian vector with distribution

$$[\mathbf{Z} \mid \mathbf{U}] \overset{\mathrm{d}}{=} \mathrm{N}_n(-\mathbf{D}^t\mathbf{\Sigma}(\mathbf{\Delta} + \mathbf{D}^t\mathbf{\Sigma}\mathbf{D})^{-1}(\mathbf{U} - \nu), \mathbf{\Sigma} - \mathbf{D}^t\mathbf{\Sigma}(\mathbf{\Delta} + \mathbf{D}^t\mathbf{\Sigma}\mathbf{D})^{-1}\mathbf{\Sigma}\mathbf{D}). \quad (3)$$

This property provides a two-step algorithm for simulating a CSN vector $\mathbf{Z}$: (i) generate samples of the Gaussian vector $\mathbf{U} \overset{\mathrm{d}}{=} \mathrm{N}_m(\nu, \mathbf{\Delta} + \mathbf{D}^t\mathbf{\Sigma}\mathbf{D})$ such that $\mathbf{U} \leq \mathbf{0}$; (ii) generate the Gaussian vector $[\mathbf{Z} \mid \mathbf{U}]$ according to (3). Generating a vector $\mathbf{U}$ conditional on $\mathbf{U} \leq \mathbf{0}$ is not direct. In particular direct seqential simulations cannot be used to generate such a vector. MCMC methods must be used instead. Here, we used a Gibbs sampling technique to simulate the vector $\mathbf{U} \mid \mathbf{U} \leq \mathbf{0}$.

The moment generating function (mgf) of a closed-skew normal density is equal to (Domínguez-Molina et al., 2003):

[redacted]

The mgf of a CSN random vector is thus the product of the usual mgf of a Gaussian vector with mean μ and covariance matrix $\mathbf{\Sigma}$ by a the m dimensional normal cpf with mean ν and covariance matrix $\mathbf{\Delta} + \mathbf{D}^t\mathbf{\Sigma}\mathbf{D}$. It is well known that even for moderate dimensions for m, the cpf Φ_m is difficult to compute.

3 Spatial skewed Gaussian processes

Let $\{Z(x)\}$ with $x \in \mathbb{R}^2$, be a spatial, ergodic, stationary, zero-mean Gaussian process with variogram

$$2\gamma(h) = \mathrm{Var}(Z(x+h) - Z(x)), \text{ for any } h \in \mathbb{R}^2$$

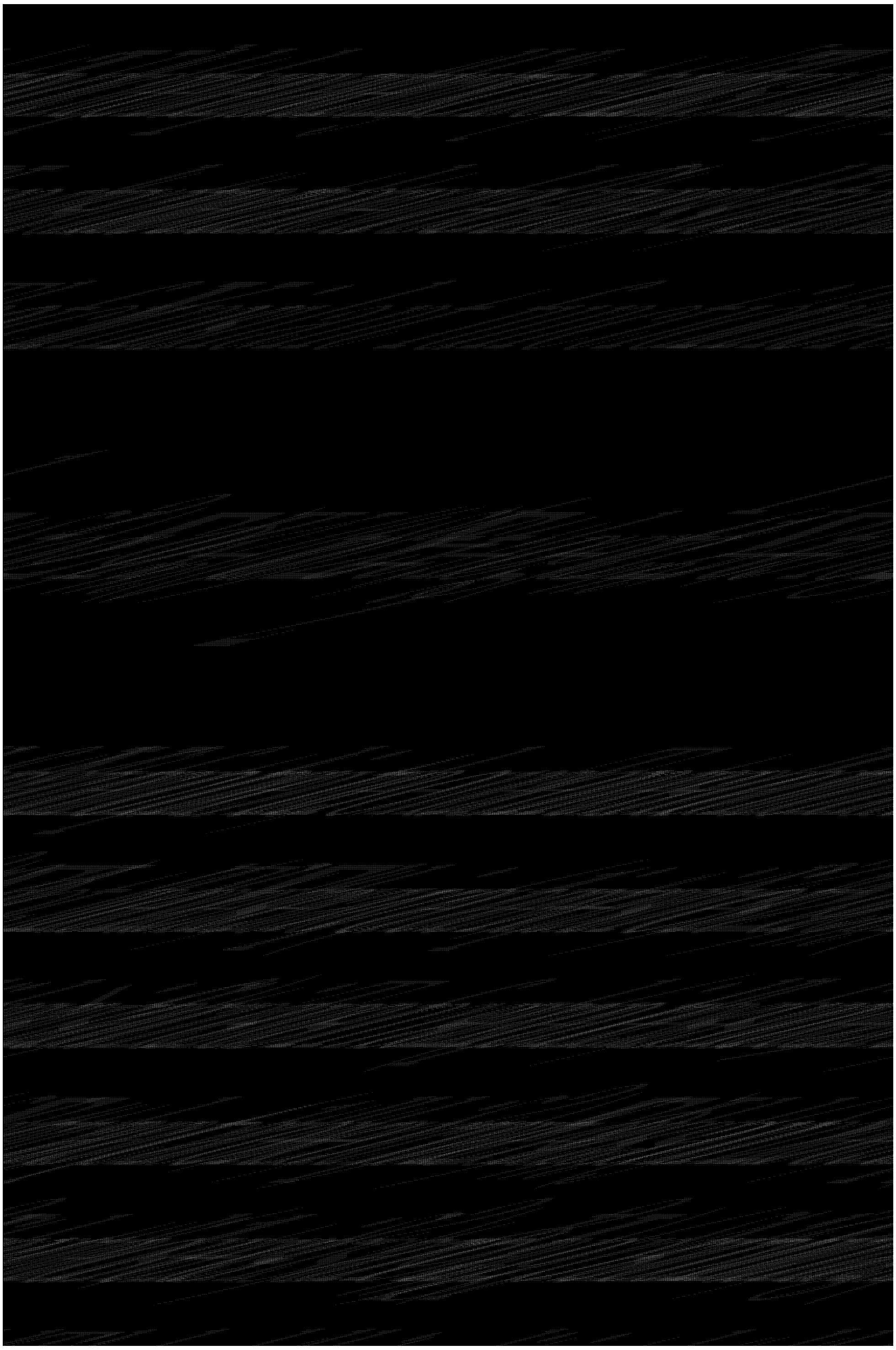
and variance $\sigma^2 = \mathrm{Var}(Z(x))$. For more details on the variogram, we refer to the following books: Wackernagel, 2003, Chilès and Delfiner, 1999, Stein, 1999 and Cressie, 1993. The covariance matrix of the random vector $\mathbf{Z} = (Z(x_1), ..., Z(x_n))^t$ built from the covariance function $c(h) = \sigma^2 - \gamma(h)$ is denoted by $\mathbf{\Sigma}$. To link this spatial structure with skew normal distributions, we simply plug the covariance matrix $\mathbf{\Sigma}$ in Equation (2). Hence, we assume in the rest of this paper that the vector $\mathbf{Z}$ is the same that the one used in Equation (2). Consequently, the process $\{Y(x)\}$ is defined through the following equality

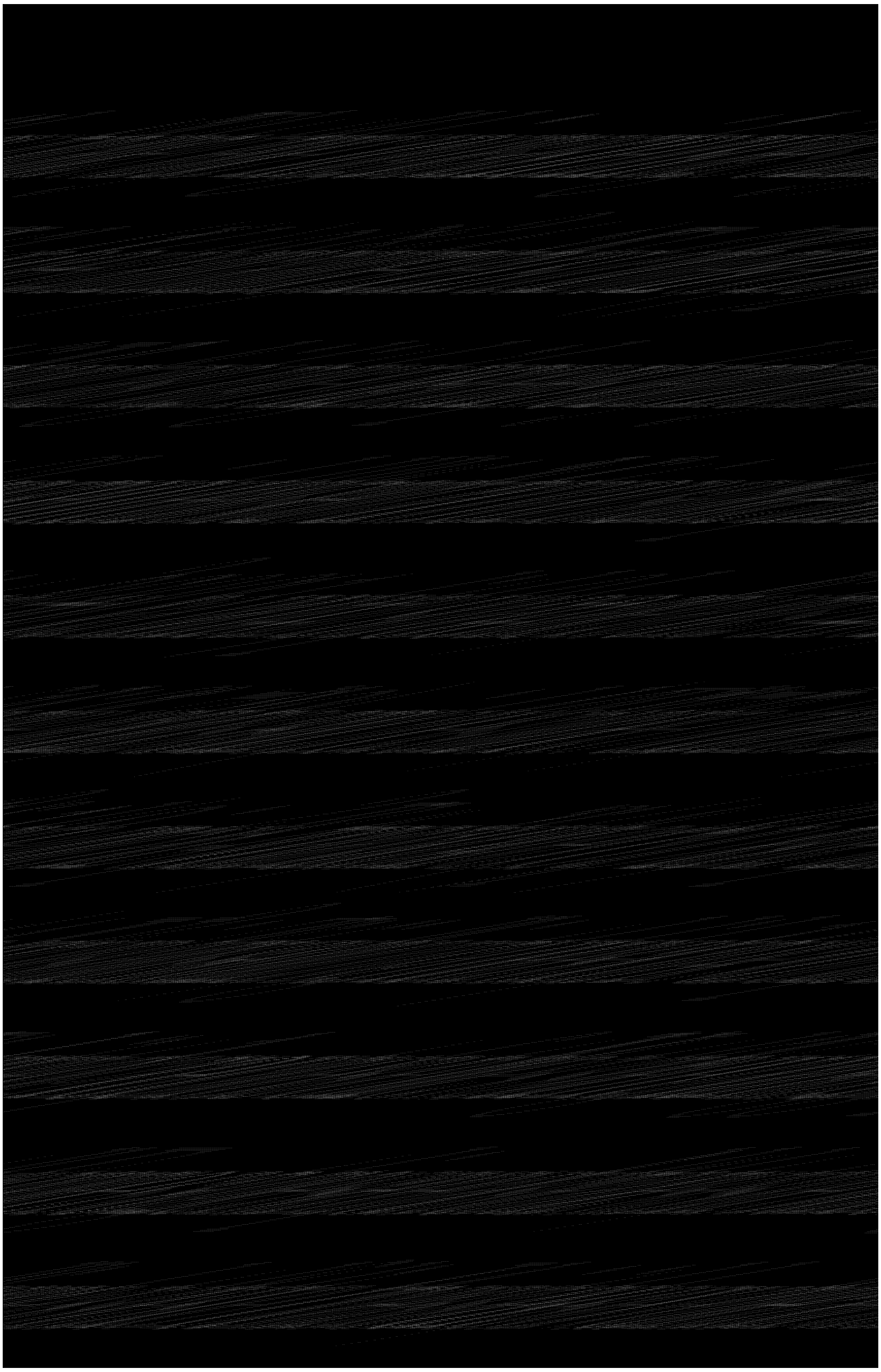
$$\mathbf{Y} \overset{\mathrm{d}}{=} \mu + [\mathbf{Z} \mid \mathbf{U} \leq \mathbf{0}].$$

This is our definition of a CSN random process. In practice, we only observe the realizations $(Y(x_1), ..., Y(x_n))^t$, but neither $\mathbf{U}$ nor $\mathbf{Z}$.

[redacted]

[redacted]





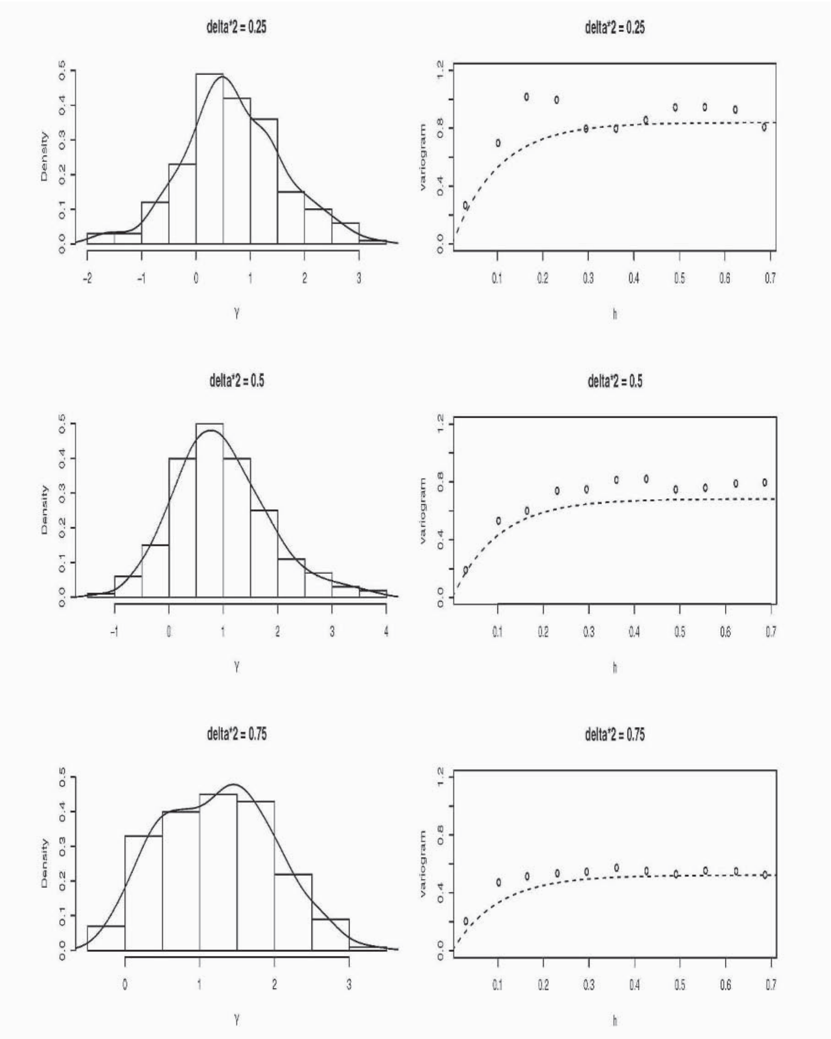


Figure 1. Histogram and variogram of simulated skewed Gaussian processes

these limitations, the variogram can be well estimated but more work is needed to estimate accurately the skewness parameter.

Finally, we believe that spatial models based on the closed-skew normal distribution can offer an interesting alternative to represent skewed data without transforming them. Still, much more research, theoretical as well as practical, has to be undertaken to determine the advantages and the limitations of such a approach.

This page intentionally blank

5 Acknowledgments

This research was supported by the National Science Foundation with the grant ATM-0327936.

References

Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scand. J. Stat.*, 12:171–178.

Azzalini, A. (1986). Further results on a class of distributions which includes the normal ones. *Statistica*, 46:199–208.

Azzalini, A. and Capitanio, A. (1999). Statistical applications of the multivariate skew normal distribution. *J. R. Statist. Soc. B*, 61:579–602.

Azzalini, A. and Dalla Valle, A. (1996). The multivariate skew-normal distribution. *Biometrika*, 83:715–726.

Chilès, J.-P. and Delfiner, P. (1999). *Geostatistics: Modeling Spatial Uncertainty*. John Wiley & Sons Inc., New York. A Wiley-Interscience Publication.

Cressie, N. A. C. (1993). *Statistics for Spatial Data*. John Wiley & Sons Inc., New York, revised reprint. Revised reprint of the 1991 edition, A Wiley-Interscience Publication.

Domínguez-Molina, J., González-Farías, G., and Gupta, A. (2003). The multivariate closed skew normal distribution. Technical Report 03-12, Department of Mathematics and Statistics, Bowling Green State University.

Domínguez-Molina, J., González-Farías, G., and Ramos-Quiroga, R. (2004). Skew-normality in stochastic frontier analysis. In Genton, M., editor, *Skew-Elliptical Distributions and Their Applications: A Journey Beyond Normality*, pages 223–241. Edited Volume, Chapman & Hall, CRC, Boca Raton, FL, USA.

Genton, M. (2004). *Skew-Elliptical Distributions and Their Applications: A Journey Beyond Normality*. Edited Volume, Chapman & Hall, CRC, Boca Raton, FL, USA.

Genton, M. and N. Loperfido, N. (2005). Generalized skew-elliptical distributions and their quadratic forms. submitted. *Annals of the Institute of Statistical Mathematics (in press)*.

González-Farías, G., Domínguez-Molina, J., and Gupta, A. (2004). The closed skew-normal distribution. In Genton, M., editor, *Skew-Elliptical Distributions and Their Applications: A Journey Beyond Normality*, pages 25–42. Edited Volume, Chapman & Hall, CRC, Boca Raton, FL, USA.

Kim, H. and Mallick, B. (2002). Analyzing spatial data using skew-gaussian processes. In Lawson, A. and Denison, D., editors, *Spatial Cluster Modelling*. Chapman & Hall, CRC.

Naveau, P., Genton, M., and shen, X. (2004). A skewed kalman filter. *Journal of Multivariate Statistics (in press)*.

Stein, M. L. (1999). *Interpolation of Spatial Data*. Springer-Verlag, New York. Some theory for Kriging.

Wackernagel, H. (2003). *Multivariate Geostatistics. An Introduction with Applications*. Springer, Heidelberg, third edition.

GRADUAL DEFORMATION OF BOOLEAN SIMULATIONS

MICKAËLE LE RAVALEC-DUPIN AND LIN YING HU
Reservoir Engineering Department,
Institut Français du Pétrole, 1 & 4 avenue de Bois Préau
92852 Rueil-Malmaison Cedex, France

Abstract. Gaussian random simulations are used in petroleum engineering and hydrology to describe permeability and porosity distributions in subsurface reservoirs. Except by luck, the generated simulations do not yield numerical flow answers consistent with the measured production data. Thus, they have to be modified, which can be done by running an optimization process. The gradual deformation method was introduced to modify Gaussian simulations. As the resulting variations are continuous, this technique is of interest for gradient-based optimizations. Based upon the gradual deformation method, a preliminary approach was suggested to modify also Boolean simulations. In this paper, we aim at going one step further. First, the gradual deformation scheme, initially developed for Gaussian probabilities, is reformulated for Poisson probabilities. It provides a new tool for varying the number of objects populating a Boolean simulation. Up to now, changing this number induced sudden object appearance or disappearance, which produced strong objective function discontinuities. Such a behavior is especially undesired when running gradient-based optimizations. Thus, we extend the proposed approach to continuously add or remove objects from Boolean simulations. The resulting algorithm integrates easily into optimization procedures and reduces, at least partially, the objective function discontinuities due to the appearance or disappearance of objects.

1 Introduction

The realizations of Gaussian random functions are often used to describe the spatial distributions of physical properties, such as permeability or porosity, in subsurface reservoirs (Journel and Huijbregts, 1978; Chilès and Delfiner, 1999). To ensure that these realizations are valuable images of a given reservoir, we have to make them consistent with all of the collected data, that is the static data and the dynamic data. Static data gather *e.g.* core data, log data, seismic data. They can be integrated in realizations using for instance kriging techniques. This subject is out the scope of this paper. Dynamic data are influenced by fluid displacements: they correspond to *e.g.* water cuts, well pressures, flow rates. These dynamic data are usually accounted for through an optimization process (Tarantola, 1987; Sun, 1994). It involves the definition and minimization of an objective function, which quantifies the discrepancy between the available dynamic data and the equivalent simulated answers. In practice, a starting realization is sequentially modified until it provides answers, which fit the required data.

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 939–948.

The gradual deformation method was initially presented as a tool to deform a Gaussian realization from a reduced number of parameters while preserving its spatial variability (Hu, 2000[a]; Le Ravalec *et al*., 2000). It is well suited for gradient-based optimizations: deforming gradually Gaussian realizations produces smooth variations in the objective function. The efficiency of this deformation technique has been pointed out in many study cases (Roggero and Hu, 1998; Le Ravalec *et al*., 2001; Ferraille *et al*., 2003).

Later on, the gradual deformation method, initially designed for Gaussian simulations, was extended to non-Gaussian simulations, more especially to Boolean or objects simulations. This type of simulation is often used to describe channel systems or fracture networks. Channels, or fractures, are recognized as objects, which populate simulations. Hu (2000[b]) suggested to apply the gradual deformation method to modify object locations. Furthermore, still based upon the gradual deformation method, Hu (2003) proposed algorithms for changing the number of objects included in a simulation. Unfortunately, this additional feature induces strong objective function discontinuities due to the sudden appearance or disappearance of objects. Such a behavior is obviously undesired for gradient-based optimizations.

In this paper, we first recap the basics of the gradual deformation method as initially presented. Then, instead of focusing on Gaussian probabilities as done up to now, we introduce a new gradual deformation formulation appropriate for Poisson probabilities. The proposed technique allows not only for gradually varying the numbers of objects populating a simulation, but also for envisioning a new Boolean simulation technique with gradual appearance and disappearance of objects. Last, a numerical experiment stresses the potential of the suggested approach.

2 Recap about the gradual deformation method

2.1 MULTI-GAUSSIAN RANDOM FUNCTION

Up to now, the gradual deformation method was intended for multi-Gaussian random functions solely. In other words, let $Y_1(\mathbf{x})$ and $Y_2(\mathbf{x})$ be two independent stationary multi-Gaussian random functions of order 2. $\mathbf{x}$ is the location. For simplicity, both functions are assumed to have zero mean and unit variance. They are characterized by identical covariances. A new random function $Y(t)$ is built by combining both functions as:

$$Y(t) = Y_1\cos(t) + Y_2\sin(t). \quad (1)$$

Whatever the t deformation parameter, Y has the same mean, variance and covariance as Y_1 and Y_2. In addition, Y is also a multi-Gaussian random function since it is the sum of two multi-Gaussian random functions. Two independent realizations y_1 and y_2 of Y_1 and Y_2 provides a continuous chain of realizations $y(t)$, which depend only on parameter t. This feature can be taken advantage of to calibrate realizations to production data. The leading idea is to investigate successively realization chains by tuning parameter t.

2.2 POISSON POINT PROCESS

The central element of Boolean simulations is a Poisson point process. It corresponds to the intuitive idea of points randomly distributed in space.

Let us consider a basic Boolean simulation populated by identical objects randomly and uniformly distributed in space. Their locations **x** comply with a Poisson point process of constant intensity. In other words, n independent uniform numbers give the location of an object in the n dimensional domain $[0,1]^n$.

A technique (Hu, 2000[b]) imagined to continuously move objects boils down to applying the gradual deformation method to object locations. However, you can not use at once the gradual deformation method for uniform numbers, because the sum of two independent uniform distributions is not a uniform distribution, but a triangular one. Thus, the uniform numbers are first turned into Gaussian numbers:

$$\mathbf{Y} = G^{-1}(\mathbf{x}) \tag{2}$$

G is the standard Gaussian cumulative distribution function. **x** is the location vector: it consists of uniform numbers. Thus, vector **Y** gathers Gaussian numbers. Let $\mathbf{x}_1$ the starting location of a given object and $\mathbf{x}_2$ another randomly and independently drawn location. Following Equation (1), a trajectory is defined from the gradual combination of the two locations:

$$\mathbf{x}(t) = G[G^{-1}(\mathbf{x}_1)\cos(t)+G^{-1}(\mathbf{x}_2)\sin(t)] \tag{3}$$

Again, $\mathbf{x}(t)$ is a uniform point in $[0,1]^n$, whatever parameter t. A two-dimensional example is reported in Figure 1.

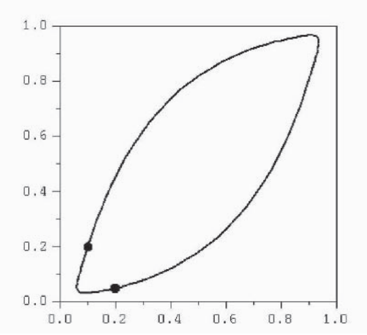


Figure 1. Trajectory defined from the gradual combination of two points in $[0,1]^2$.

2.3 NUMBER OF OBJECTS

In 1D, Poisson points delimit intervals whose lengths are independent random variables with an exponential distribution. Hu (2003) came back to this property to gradually modify the number of objects in a simulation. First, this author transformed the interval lengths to get Gaussian numbers. Then, he used the gradual deformation method to modify the Gaussian numbers as explained above, before applying the inverse transformation. This procedure allows for gradually deforming the interval lengths, which results in a variation in the number of objects.

3 Reformulating the gradual deformation method for Poisson probabilities

The technique proposed above for changing the number of objects induces strong variations in the objective function, because of the sudden appearance or disappearance of objects. Such a behavior can not be accommodated by gradient-based optimization processes. In this section, we propose a novel approach to alleviate this undesired feature. Instead of going back to the basic gradual deformation relation (Equation 1), which is appropriate only for Gaussian probabilities, we focus on a new formulation for Poisson probabilities.

3.1 POISSON NUMBERS

Our purpose is to modify gradually the number of objects populating a Boolean simulation. This number is random and respects a Poisson probability law. The probability for this non negative number to be *n* is given by:

$$\Pr(N = n) = \exp(-\lambda)\frac{\lambda^n}{n!} \tag{4}$$

where *N* is a Poisson variable with parameter λ. It can be shown that its mean and variance equal λ.

To produce *n*, we generate a 1D Poisson point process with unit mean in an interval of length λ. As shown in Figure 2, we sequentially draw independent intervals (OE_1, E_1E_2, …, E_nE_{n+1}) from an exponential distribution of unit mean, denoted γ_1. We add the intervals until their sum is more than λ.

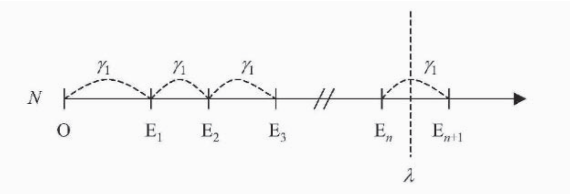


Figure 2. Simulating realization *n* from a Poisson variable with parameter λ.

The cumulative distribution function of an exponential variable with unit mean is:

$$F(x) = 1\text{-}\exp(\text{-}x). \tag{5}$$

Let *r* be a uniform deviate drawn between 0 and 1. Thus, 1-*r* is also a uniform deviate lying within the same range. We assume that 1-*r* = 1-exp(-*x*). To produce the successive intervals, we repeat x = -Log(r) for successive r values. Realization n of Poisson variable N is the biggest integer so that:

$$\sum_{i=1}^{n} - \mathrm{Log}(r_i) < \lambda \,. \qquad (6)$$

3.2 GRADUAL DEFORMATION OF POISSON PROBABILITIES

The gradual deformation method as presented in Section 2 applies to Gaussian random variables, which continuously vary in $\Re$. As Poisson variables provide integers, a new deformation scheme is necessary.

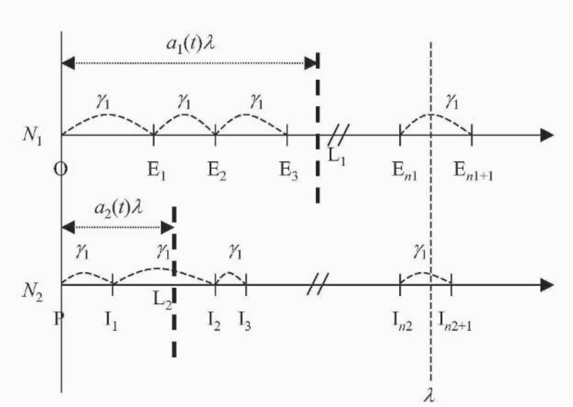


***Figure* 3.** Gradual deformation of a Poisson variable with parameter λ by combining two independent Poisson variables, N_1 and N_2.

The sum of two independent Poisson variables with parameters λ and μ, respectively, is also a Poisson variable with parameter $\lambda + \mu$. This fundamental property is the key point of the new deformation algorithm developed in this paper. Let N_1 and N_2 be two independent Poisson variables, both with parameter λ (Figure 3). We suggest to gradually deform the parameters of the two added Poisson variables, while respecting the following constraint: the sum of the two deformed parameters equals λ. To avoid any confusion, we will refer to parameter for Poisson variables and to deformation coefficient for gradual deformation. Thus, a new Poisson variable with parameter λ is obtained from:

$$N(t)\{\lambda\} = N_1\{a_1(t)\lambda\} + N_2\{a_2(t)\lambda\} \qquad (7)$$

with $a_1(t)$+ $a_2(t)$=1. $a_1(t)\lambda$ and $a_2(t)$ λ are the parameters of N_1 and N_2, respectively. t is the gradual deformation coefficient. Because of the periodicity of the trigonometric functions, we choose $a_1(t)$ = cos^2(t) and $a_2(t)$ = sin^2(t). In addition, the extension of this parameterization to the combination of more than two Poisson variables is straightforward.

Variations in the deformation coefficient induce variations in the parameters of the added Poisson variables, which impact the realizations of these variables. As a result, their sum is also changed. Starting from two realizations n_1 and n_2 of N_1 and N_2, a chain of realization $n(t)$ is built by varying deformation coefficient t. When t is 0, n equals n_1; when t is $\pi/2$, n equals n_2. This deformation process is depicted in Figure 4 for a Poisson variable with a parameter of 10.

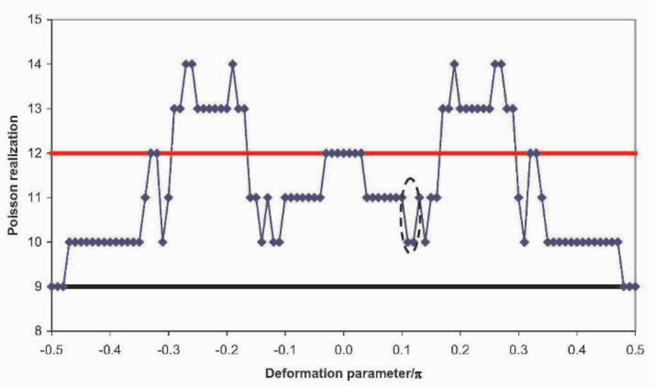


Figure 4. Gradual deformation of the realization of a Poisson variable with a parameter of 10. When t is 0, n is the same as the starting realization (12). When t is $\pi/2$, n is the same as the second realization (9). We will focus on the 4 surrounded dots in the subsequent sections.

Another possibility to gradually deform Poisson variables would be to move a segment of constant length λ along an infinite axis populated with intervals drawn from an exponential distribution. The number of complete intervals included in the segment is a realization of a Poisson variable with parameter λ. It changes depending on the location of the segment.

3.3 BUILDING SUCCESSIVE CHAINS

In the case of multi-Gaussian variables, the gradual deformation method judiciously integrates optimization procedures. Briefly, chains of realizations are successively investigated until an appropriate realization is identified. The chains are built from the "optimal" realization determined for the previous chain and a second randomly drawn realization.

The procedure for Poisson variables is very similar. The gradual deformation of two Poisson variables (a starting one plus a second independent one) as explained above yields a first realization chain. The investigation of this chain provides a first "optimal" realization (*i.e.*, segments), which minimizes the objective function. It is used as the starting realization of the following chain. To fully define the second chain, an independent Poisson realization is randomly drawn. Again, we can explore this second chain and try to identify a realization, which reduces further the objective function. This process is repeated until the objective function is small enough.

4 Smooth appearance and disappearance of Boolean objects

The gradual deformation of Poisson variables as introduced in the previous section allows for varying the number of objects in a Boolean simulation, but it still involves sudden appearance or disappearance of objects. This feature results in strong objective function variations, which cannot be accommodated by gradient-based optimizations. We extend the gradual deformation principles developed for Poisson variables one step further and suggest a novel technique to generate Boolean simulations with continuous appearance or disappearance of objects.

4.1 PRINCIPLE

As shown in Figure 2, we generate a realization n of a Poisson variable with parameter λ by adding intervals until the sum of their lengths is more than λ. n is the integer so that $OE_n < \lambda$ and $OE_{n+1} > \lambda$.

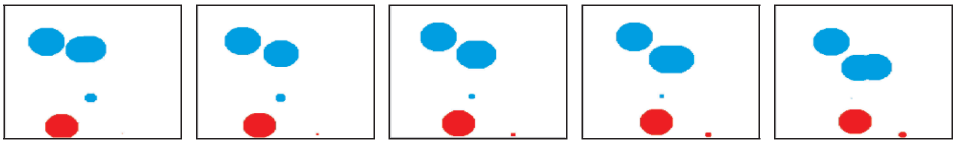


Figure 5. Gradual deformation of the number and locations of objects. Objects appear and disappear continuously. Two kinds of objects are identified depending on the Poisson variable they refer to (N_1 or N_2). In this case, a grey ellipse gets smaller and eventually vanishes while a black one appears and expands.

Let L_1 be the point so that OL_1 equals the parameter of Poisson variable N_1 (Figure 3). It moves as the deformation coefficient varies. When t is 0, the parameter of Poisson variable N_1 is λ. At this point, OL_1 contains n complete intervals plus a truncated one. We consider that the Boolean simulation is populated by n+1 objects. We assume that their sizes are derived from an anamorphosis function applied to the lengths of the n complete intervals and of the truncated one. If the deformation coefficient increases, the length of OL_1 decreases, which also induces a decrease in the length of the interval $E_{n1}L_1$. In other words, the size of the $(n+1)^{\text{th}}$ object decreases: this objects continuously vanishes. If the parameter of N_1 decreases further, at some point the n^{th} object will also gets smaller and smaller. In parallel, the parameter of N_2 increases. When t is 0, the parameter of this Poisson variable is 0. Then, an object starts to appear. Its size depends on the length of the increasing interval PL_2 (Figure 3), L_2 being the point so that PL_2 equals the parameter of Poisson variable N_2. When $PL_2 = PI_1$ (Figure 3), the first object is complete. If the parameter of N_2 increases again, a second object appears. Its size depends on the length of I_1L_2. The proposed method is illustrated by the example shown in Figure 5.

4.2 NUMERICAL EXPERIMENT

We consider a synthetic reservoir model (Figure 6) formed of lenses with a permeability of 50 mD, embedded in a reservoir rock with a permeability of 500 mD. For the sake of simplicity, porosity is assumed to equal 30% everywhere. The reservoir model is discretized over a regular grid of 200x200 gridblocks. The size of a gridblock is given by DX = 1 m and DY = 0.8 m. At the beginning, the reservoir is fully saturated by oil. Then, water has been injected at well I at 100 m^3/day for 100 days while oil has been produced at well P at constant pressure. The numerical simulations are run using 3DSL (Batycky *et al.*, 1997). The simulated pressures at well I and fractional water flow at well P are depicted in Figure 6

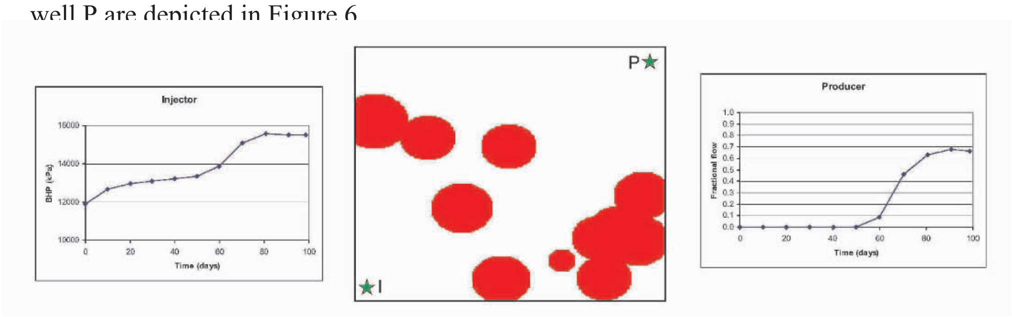


Figure 6. Synthetic reservoir model and corresponding pressures at the injecting well and fractional water flow at the producing well.

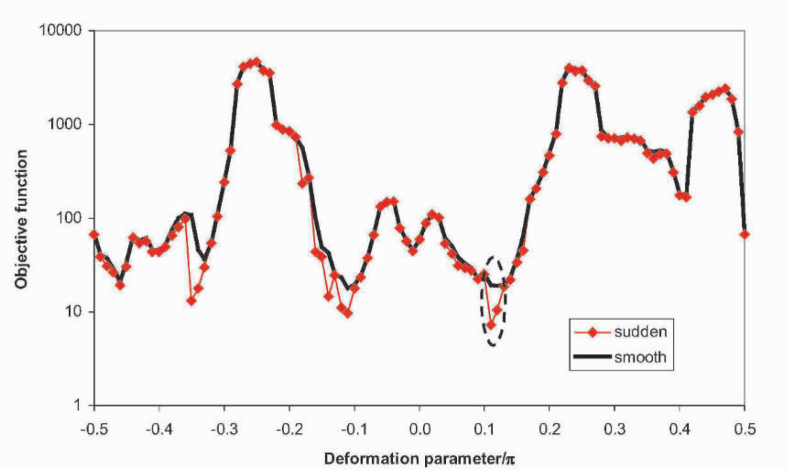


Figure 7. Objective function against the deformation coefficient. Thick curve: smooth appearances and disappearances of objects. Thin curve with diamonds: sudden appearance and disappearance of objects.

At this point, we assume that the synthetic reservoir model is unknown. The only available data are the pressures and the fractional flow shown above (Figure 6) and some prior geological information stating that the number of lenses can be approximated by a Poisson variable with parameter 10. First, we generate an initial guess for the synthetic reservoir model. Then, by applying the gradual deformation process introduced in Sections 3 and 4.1, we build a chain of Boolean realizations. In this example, the deformation coefficient impacts the number of lenses and their locations. The number of complete objects is the one displayed in Figure 4. For all of the realizations, we simulate the previously described water injection test. We compute the objective function, which measures the mismatch between the reference data (Figure 6) and the simulated ones for the explored Boolean realizations (Figure 7). Two distinct cases are envisioned. First, the lenses appear and disappear suddenly. Second, they appear and disappear smoothly. In the first case, the objective function exhibits sudden jumps, which are reduced in the second case.

Let us focus on the four surrounded points in Figure 7. In the sudden appearance and disappearance case, the objective function shows a very discontinuous variation. On the contrary, it varies continuously when objects appear and disappear smoothly. The four corresponding realizations are presented in Figures 8 and 9. In Figure 8.2, the sudden disappearance of a lense enlarges a lot the flow path towards the producing well. In Figure 9.2, the lense does not vanish: it gets smaller. Thus, the flow path does not open as much as in Figure 8.2. In Figure 9.4, the lense is so small that it does no longer affect flow: the objective functions are the same for Figures 8.4 and 9.4.

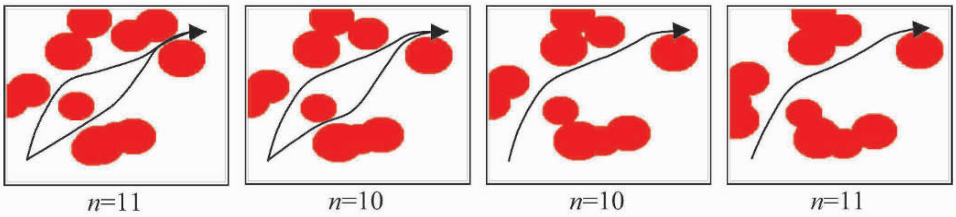


Figure 8. Gradual deformation of the number and location of objects with sudden appearance and disappearance. These 4 realizations correspond to the 4 surrounded points in Figure 7.

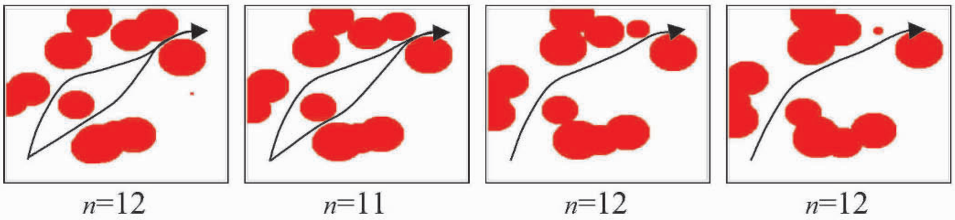


Figure 9. Gradual deformation of the number and location of objects with smooth appearance and disappearance. These 4 realizations correspond to the 4 surrounded points in Figure 7.

As explained at the end of Section 3, many chains could be investigated successively. The only difference with Section 3 is that we also account for the truncated segments.

5 Conclusions

The following main conclusions can be drawn from this study.

- We developed a new gradual deformation process, which is appropriate for Poisson variables. It allows for varying the number of objects populating a Boolean realization.

- We suggested to associate the sizes of the objects to the Poisson point process used to generate the number of objects. This additional feature makes it possible to smoothly introduce or remove objects from the Boolean realization all along the gradual deformation process. The suggested algorithm integrates easily into optimization procedures. It reduces significantly the objective function discontinuities due to the sudden appearance or disappearance of objects and is well suited for gradient-based optimizations.

- However, the suggested method does not prevent the objective function from being discontinuous. In some cases, the displacement of an object from a single grid block can drastically modify the connectivity of the realization and produce strong objective function fluctuations. The proposed gradual deformation technique with smooth appearance and disappearance of objects does not eliminate such discontinuities, but it contributes to reduce their effects.

Acknowledgements

We would like to acknowledge the industry sponsors (Groupement Berkine, BHP-Billiton, Eni-Agip, Gaz de France, Petrobras and Total) of CONDOR II for supporting this research.

References

Batycky, R.P., Blunt, M.J., and Thiele, M.R., A 3D field-scale streamline based reservoir simulator, SPERE, 4, November 1997.

Chilès, J. P., and Delfiner, P., Geostatistics: Modeling spatial uncertainty, Wiley, New York, 695p, 1999.

Feraille, M., Roggero, F., Manceau, E., Hu, L.Y., Zabalza-Mezghani, I., and Costa Reis, L., Application of advanced history matching techniques to an integrated field case study, Annual Technical Conference & Exhibition, SPE 84463, Denver, USA, 5 - 8 October, 2003.

Hu, L.-Y., Gradual deformation and iterative calibration of Gaussian-related stochastic models, *Math. Geol*., 32(1), 2000^a.

Hu, L.-Y., Geostats 2000 Cape Town, WJ Kleingeld and DG Krige (eds.), 1, 94-103, 2000^b.

Hu, L.-Y., History matching of object-based stochastic reservoir models, SPE 81503, 2003.

Journel, A., and Huijbregts, C.J., Mining Geostatistics, Academic Press, London, 600p, 1978.

Le Ravalec, M., Noetinger, B., and Hu, L.Y., The FFT moving average (FFT-MA) generator: an efficient numerical method for generating and conditioning Gaussian simulations, *Math. Geol*., 32(6), 701-723, 2000.

Le Ravalec, M., Hu, L.-Y. and Noetinger, B., Stochastic Reservoir modeling constrained to dynamic data: local calibration and inference of the structural parameters, *SPE Journal*, 25-31, March 2001.

Roggero, F., and Hu, L.-Y., Gradual deformation of continuous geostatistical models for history matching, SPE ATCE, 49004, New Orleans, LA, USA, 1998.

Sun, N.-Z., Inverse problems in groundwater modeling, Kluwer Academic Publishers, The Netherlands, 1994.

Tarantola, A., Inverse problem theory – Methods for data fitting and model parameter estimation, Elsevier Science Publishers, Amsterdam, Netherlands, 1987.

WHEN CAN SHAPE AND SCALE PARAMETERS OF A 3D VARIOGRAM BE ESTIMATED?

PÅL DAHLE, ODD KOLBJØRNSEN, and PETTER ABRAHAMSEN
Norwegian Computing Center, Box 114 Blindern, NO-0314 Oslo, Norway

Abstract. We have used a method of least squares to fit full 3D variogram models to data, and have tested it on data taken from Gaussian random fields. The empirical variogram estimates are made using various lag grid definitions and the best of these grids is identified. Our results suggest that some 200 vertical wells are needed for obtaining reliable estimates of the azimuth and dip anisotropy angles, while some 50 wells seem sufficient for the horizontal ranges and the sill. For the vertical range 10 wells are sufficient.

1 Introduction

Estimation of a variogram is an important issue in spatial problems because inference regarding spatial variables often rest on a variogram model. A common approach is to fit the model variogram "by eye". Although this approach is convenient in one dimension, it becomes intractable in two and three dimensions. Unless one wants to let the variogram model be based on knowledge about the geology of the field in question or similar fields, this implies that an algorithm for automatic fitting has to be implemented.

We have done a large computer study for automatic fitting of variograms in three dimensions; similar studies in one dimension are found in Webster and Oliver (1992), Pardo-Igúzquiza (1999), and Chen and Jiao (2001).

Although automatic fitting in one dimension may provide valuable information about the variogram, all directions should be treated simultaneously, as variance contributions identified for one direction should also hold for the other directions, as pointed out by Gringarten and Deutsch (2001).

Although it is important to treat all directions simultaneously, all variogram parameters should not necessarily be optimised. Unless sufficient data is available, an automatic fit may potentially lead to variograms that are non-geological. The dip angle, for instance, requires a large number of observations for reliable estimation, and if too few observations are available, an opposite dip may easily be the result. In such cases, one is better off using a qualified guess.

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 949–958.

2 Fitting model variograms to data

2.1 ESTIMATING EMPIRICAL VARIOGRAMS

A commonly used estimator of the variograms is the method of moments estimator (Matheron, 1963). Given a random variable $Z(\mathbf{x})$, the estimator $\hat{\gamma}(\mathbf{h})$ is

$$\hat{\gamma}(\mathbf{h}) = \frac{1}{2N_{\mathbf{h}}} \sum_{N(\mathbf{h})} \left(Z(\mathbf{x}_i) - Z(\mathbf{x}_j)\right)^2, \qquad \mathbf{h} \in \mathbb{R}^3 \tag{1}$$

where $Z(\mathbf{x}_i)$ is the value of the random variable Z at some point $\mathbf{x}_i$, and $\mathbf{h}$ is a vector connecting the two points $\mathbf{x}_i$ and $\mathbf{x}_j$. $N(\mathbf{h})$ denotes all pairs $\{(\mathbf{x}_i, \mathbf{x}_j)\}$ that may be connected by the lag vector $\mathbf{h}$, and $N_{\mathbf{h}}$ is the cardinality of $N(\mathbf{h})$.

The variogram estimator proposed by Matheron is sensitive to outliers. Hawkins and Cressie (1984) recognised this and proposed a robustified alternative

$$\bar{\gamma}(\mathbf{h}) = \frac{1}{2g(N_{\mathbf{h}})} \left\{ \sum_{N(\mathbf{h})} |Z(\mathbf{x}_i) - Z(\mathbf{x}_j)|^{\frac{1}{2}} \right\}^4, \qquad \mathbf{h} \in \mathbb{R}^3 \tag{2}$$

where $g(N_{\mathbf{h}})$ is a function of the cardinality: $g(N_{\mathbf{h}}) = N_{\mathbf{h}}^4\,(0.457 + 0.494/N_{\mathbf{h}})$.

2.2 PARAMETRIC MODEL VARIOGRAMS

The parametric model variograms that are to be fitted to the empirical estimates may be denoted $\gamma(\mathbf{h};\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ are variogram parameters like ranges, anisotropy angles, and sill. For some variogram models there are also additional parameters.

In this paper we have explored the four variogram types: spherical, exponential, general exponential (stable), and Gaussian. The mathematical forms of these variograms may be found in, for example, Yaglom (1987).

2.3 DISTANCE ESTIMATORS

In order to optimise the parameters of the model variogram, we must minimise some distance measure between the parametric and empirical variograms. Alternatively, we can also use a maximum likelihood estimation, but such an approach is computationally much more demanding and have not been considered.

As a distance measure, we have used the *approximated weighed least squared* estimator of Cressie (1985):

$$\sum_{i=1}^{n} N_{\mathbf{h}_i} \left(\frac{\gamma^*(\mathbf{h}_i)}{\gamma(\mathbf{h}_i;\boldsymbol{\theta})} - 1 \right)^2, \tag{3}$$

where $\mathbf{h}_i, i = 1, 2, \ldots, n$ are the lags for which the variogram model $\gamma(\mathbf{h}_i;\boldsymbol{\theta})$ and the empirical variogram estimator $\gamma^*(\mathbf{h}_i)$ are to be compared. The $\boldsymbol{\theta}$ are the parameters to be optimised.

To increase the amount of data behind each variogram estimate, lag vectors are collected in bins represented by the grid cells of a lag grid. In this grid, the vector

$\mathbf{h}_i$ extends to the centre of the ith grid cell and represents all lag vectors in this cell. The variogram estimate $\gamma^*(\mathbf{h}_i)$, thus becomes an average of the variogram estimates for all lag vectors belonging to the ith cell.

Since the distance estimator in (3) is nonlinear in $\boldsymbol{\theta}$, nonlinear optimisation techniques must be employed. We have used a Gauss–Newton type optimisation (see for example Gill (1981)) in which the distance estimator is approximated to second order and the Hessian matrix is approximated using the first derivative Jacobian matrix. The Jacobian matrix, which is also used to compute the gradient, was calculated using numerical derivatives.

3 Model setup

3.1 DATA SET

The data set was obtained from Gaussian random fields generated with known variograms. Four variogram types were used: spherical, exponential, Gaussian, and general exponential, and the variogram model parameters were

1. range	2. range	3. range	azimuth	dip	sill	α
200	100	10	60°	3°	1	1.5

where the parameter α is the exponent in the general exponential variogram, and does not apply to the other variograms. The 1. and 2. ranges are predominantly horizontal while the 3. range is predominantly vertical.

The fields were generated in a 500m $\times$ 500m $\times$ 20m cube using a grid consisting of $100 \times 100 \times 50$ cells.

For each variogram type, 100 stochastic realisations of the Gaussian random field were drawn, and for each realisation, data from 10, 50, 100, and 200 randomly chosen vertical wells were collected. This gives us a total of 1600 different data sets containing either 500, 2500, 5000, or 10 000 observations.

3.2 ESTIMATING EMPIRICAL VARIOGRAMS

The empirical variogram estimates were made in a regular grid of lag vectors. Such a lag grid is fully specified once we have given the size of the grid and the number of grid cells for each Cartesian direction.

If the grid definition is altered the empirical variogram estimates change, and this, in turn, changes the shape and scale parameters of the optimised model variograms. To investigate how sensitive these parameters are to a grid change, we calculated empirical variograms using 18 different grid definitions. These grids were made by combining the three lag grid sizes $(500\text{m}, 500\text{m}, 20\text{m})$, $(300\text{m}, 300\text{m}, 10\text{m})$, and $(100\text{m}, 100\text{m}, 4\text{m})$, which give the size of the lag grid for each Cartesian direction, with six different types of grid binning: $(161 \times 161 \times 81)$, $(81 \times 81 \times 41)$, $(41 \times 41 \times 21)$, $(21 \times 21 \times 11)$, $(11 \times 11 \times 5)$, and $(5 \times 5 \times 3)$. Lag vectors extending outside the chosen lag grid were not included in the variogram estimate.

All variogram estimates were done using both the traditional and the robust estimators given in section 2.1. Thus, for each data set 36 different empirical variogram estimates were made. Since there were 1600 different data sets, a total of 57 600 variogram estimates were made.

3.3 THE FITTING OF PARAMETRIC MODEL VARIOGRAMS

When the shape and scale parameters of the model variograms were estimated, we initially assumed that the variogram model was known, that is, the model to be fitted had the same variogram as the variogram model producing the data. When the lag grid definition that gave the best parameter estimates had been identified, however, we also made successful attempts to identify the correct variogram model by fitting different models against the data and comparing residuals.

If good estimates of one or more parameters are known prior to the model fitting, these parameters may be held constant during the optimisation. The sill, for instance, is easily estimated directly from data, and this estimate is likely to be better than the estimate obtained from a multi-dimensional optimisation. Fixing the sill during the optimisation, however, will affect the ranges and anisotropy angles as well. To see which of the two approaches gives the overall best parameter estimates, all fits were made twice; first with a pre-calculated, empirical sill and then with a freely optimised sill. This gave a total of 115 200 fits.

3.4 EVALUATING THE QUALITY OF THE FITS

To evaluate the quality of the optimised model parameters we used a root-mean-square error (RMSE) measure

$$\text{RMSE} = \sqrt{\mathrm{E}\left\{(\hat{\theta} - \theta)^2\right\}} \tag{4}$$

where θ is the true parameter value (as specified in the variogram of the Gaussian random field) and $\hat{\theta}$ is the estimator for the parameter. The RMSE was based on parameter values obtained by fitting model variograms to data from 100 repetitions of the same Gaussian random field, and gave a measure of the total error involved in the parameter estimate.

When evaluating the quality of the estimator $\hat{\theta}$, we may make a comparison with the zero estimator ($\theta = 0$). A minimum requirement for $\hat{\theta}$ is that its RMSE is smaller than the RMSE of the zero estimator; at least for ranges and sill.

By comparing RMSE values obtained using different lag grid definitions and with either an empirical or an optimised sill, the RMSE may also be used to identify an optimal fitting strategy.

4 Results

From the 115 200 variogram fits that were made we got a total of 1152 different cases. In Figure 1, we have plotted the RMSE measures for all the model variogram

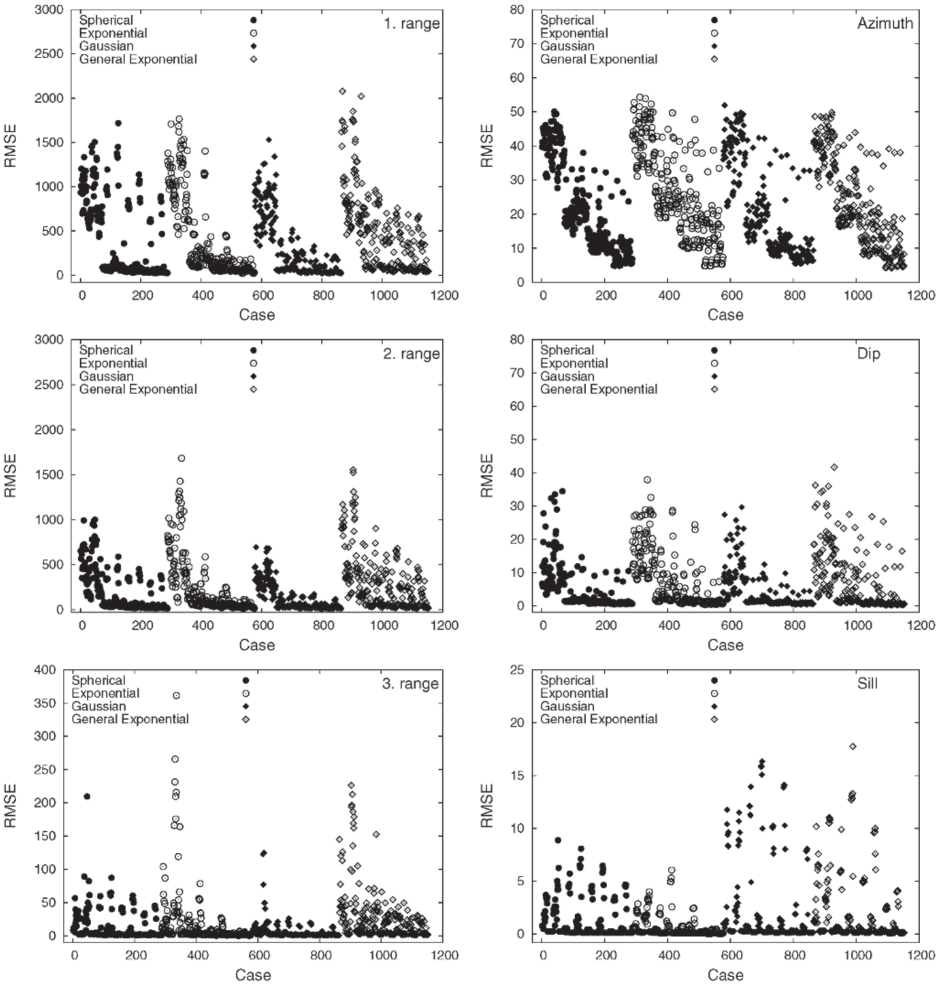


Figure 1. RMSE for variogram model parameters when different variogram models were fitted to empirical variogram estimates. Each RMSE value is based on fits to 100 realisations of the same Gaussian random field.

parameters of the different cases, except the variogram parameter of the general exponential variogram, which is discussed in section 4.3. For clarity, we have used different plotting symbols for the different variogram types. Since the variogram fit settings vary in the same manner within each variogram type, we may make direct comparisons between the different types.

First, we note that the four different variogram types seem to have fairly similar error trends, but that errors are generally somewhat higher for the general exponential variogram. Since the general exponential variogram has an extra parameter which is correlated with both ranges and sill, this is to be expected. We also note that errors tend to decrease to the right within each variogram type, due to an increased number of wells being used in the empirical variogram estimate. The "outliers" from this trend are caused by poor optimisation settings like a bad choice of lag grid or bad sill treatment.

Since the problem of finding the best optimisation settings is the same for each variogram type, we have chosen only to study the spherical variogram in more detail. Conclusions made for this variogram type are valid also for the other types.

In order to identify the factors that are important for the RMSE measures, we performed a variance analysis on the 288 error measures given for the spherical variogram type. This analysis showed that for the optimisation of the 1. and 2. ranges (horizontal), the number of wells was most important, while the treatment of the sill (optimised vs. empirical) was second most important. These two factors were also the more important for the optimisation of the 3. range (vertical), but in the reverse order. Also, for the spherical and Gaussian variograms, the lag grid size turned out to be more important than the number of wells. With the sill parameter, as with the 3. range, the most important factor was the treatment of the sill. The second most important factor was the lag grid size.

For the optimisation of the azimuth and dip angles, the number of wells was the more important factor for the RMSE measure, followed by the lag grid size.

4.1 SENSITIVITY WITH RESPECT TO MAXIMUM LAG AND WHETHER SILL IS OPTIMISED OR ESTIMATED EMPIRICALLY.

Guided by the variance analysis, we have plotted RMSE against the number of wells included in the data sets. The plots are shown in Figure 2. The RMSE have been connected into six curves corresponding to different lag grid sizes and whether the sill was optimised or estimated directly from data. Each point in these curves represents an arithmetic mean of 12 values obtained with different lag grid binning and different empirical variogram estimators.

The plots show that when the sill is optimised, the ranges and sill are sensitive to the lag grid size, and when the grid is so small that all lag vectors included in the empirical variogram estimate are considerably shorter than the true ranges, the sill and ranges become overestimated. If an empirical sill is used, however, the sensitivity is reduced to a minimum.

The anisotropy angles do not seem to be sensitive to the sill treatment, but show a clear dependency on the lag grid size. According to Figure 2, the best angle estimates are obtained when the smallest lag grid is used for the empirical variogram estimates. This is to be expected as anisotropy is more pronounced in regions where the correlation is strong. Outside the range, for instance, there is no information about anisotropy at all, and for the larger lag grids, we are therefore including random noise in the variogram estimate.

Based on the results presented in Figure 2, we conclude that the sill should be estimated directly from data and not optimised along with the rest of the model parameters. In the following, we concentrate on such variogram fits. Moreover, we shall only use the smallest lag grid, as this leads to the best anisotropy angle estimates.

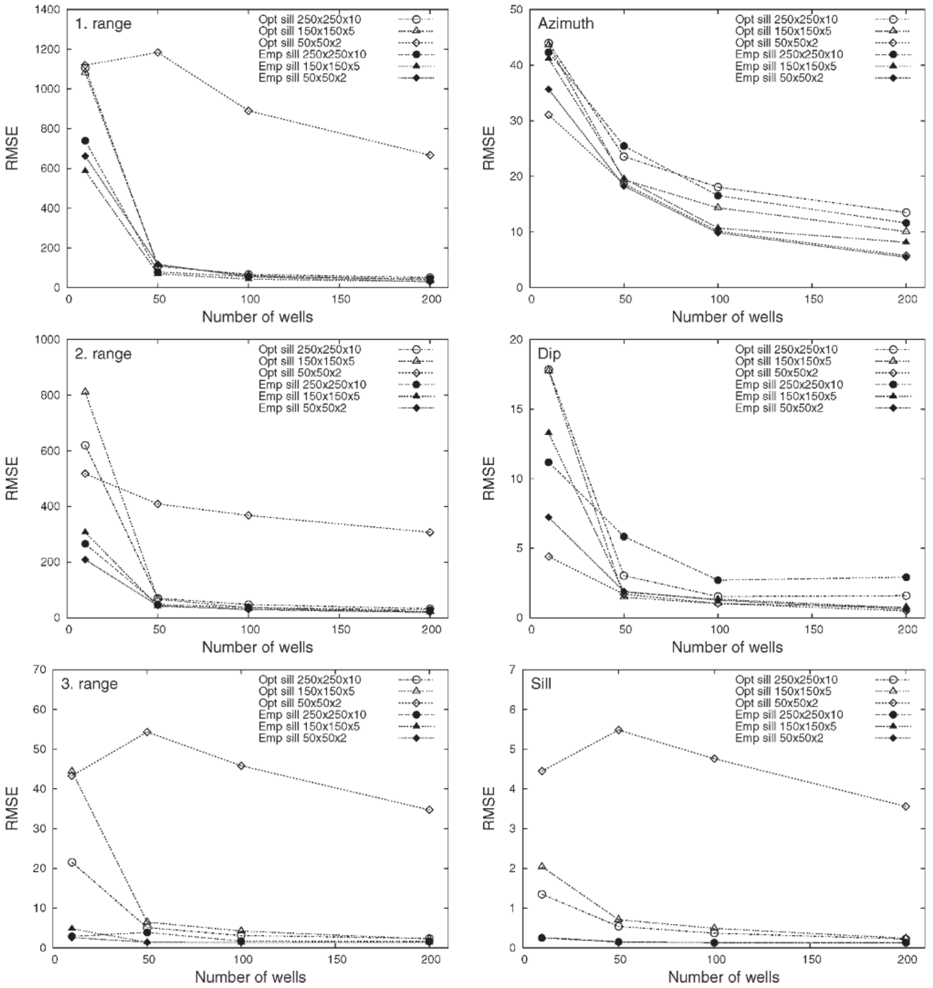


Figure 2. RMSE plotted against the number of wells. The curves correspond to different maximum lags and whether the sill was optimised or estimated directly from data.

4.2 SENSITIVITY WITH RESPECT TO CELL SIZE

To investigate how the grid binning affects the model parameter estimates, we have plotted RMSE against grid mesh in Figure 3. Each point represents an arithmetic mean of the RMSE values obtained with the two variogram estimators. Four curves are given, corresponding to different numbers of wells. For the 1. and 2. ranges, however, the curves corresponding to 10 wells are out of scale.

Figure 3 shows that the parameter optimisation is rather insensitive to the grid mesh and that essentially the same RMSE measures are obtained, unless very coarse-meshed grids are used. The poor performance for the coarsest grid is related

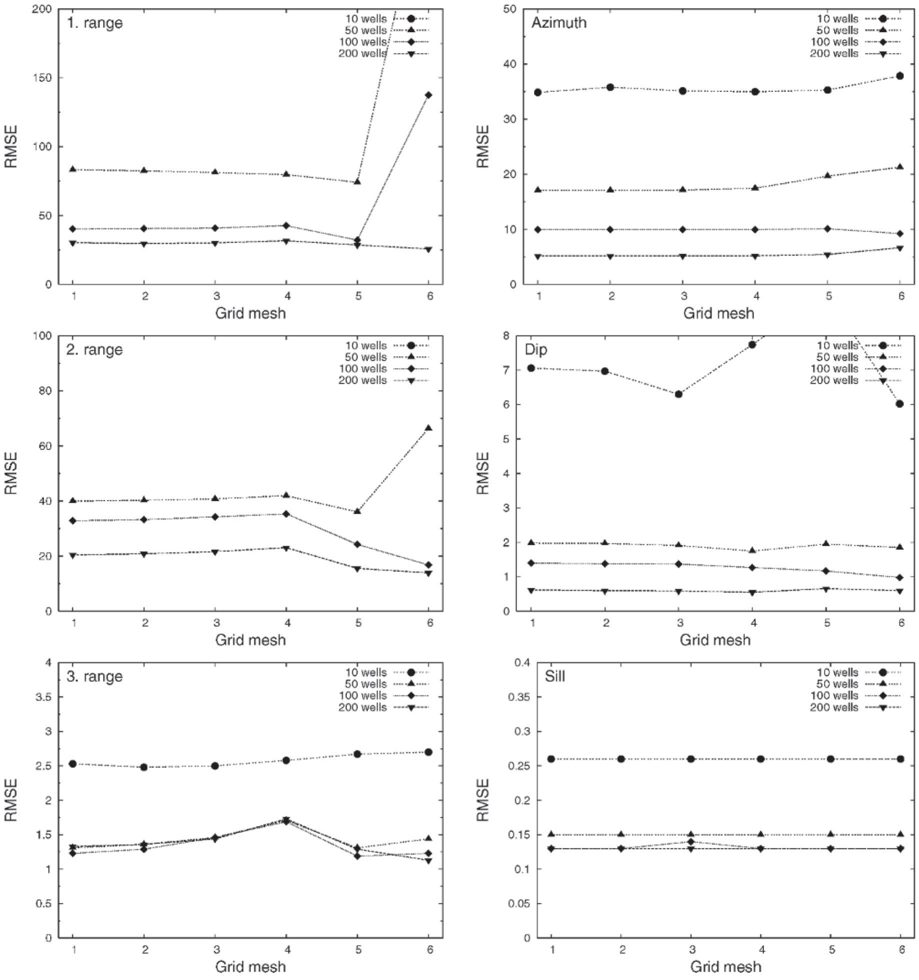


Figure 3. RMSE plotted against grid mesh. Unit 1 corresponds to the most fine-meshed grid (161×161×81) and unit 6 to the coarsest (5×5×3).

to the extensive smoothing, and such grids should be avoided. The finest-meshed grids, on the other hand, give slow optimisations, and the grid having $41 \times 41 \times 21$ cells are therefore chosen as the preferred one.

4.3 GENERAL EXP. VARIOGRAM FITS: THE VARIOGRAM PARAMETER

The RMSE plots for the variogram parameter of the general exponential variogram are similar to those given in Figures 2 and 3 and are therefore not given. Again, it is found that the best parameter estimates are obtained for the smallest lag grid. Somewhat surprisingly, however, it is found that slightly better estimates are obtained when the sill is optimised rather than estimated from data.

Table 1. The variogram types that best fit data from 100 realisations of the four Gaussian random fields. Each table entry gives the number of realisations for which the particular variogram model fitted data best.

		Variogram of Gaussian random field											
		200 wells				50 wells				10 wells			
		Sph	Exp	Gen	Gau	Sph	Exp	Gen	Gau	Sph	Exp	Gen	Gau
Best model	Sph	97	20	–	–	94	64	1	–	69	37	4	–
	Exp	3	80	–	–	5	36	2	–	28	61	1	–
	Gen	–	–	100	–	1	–	96	–	3	1	96	3
	Gau	–	–	–	100	–	–	1	100	–	1	3	97

4.4 FINDING THE VARIOGRAM MODEL THAT BEST FITS DATA

Using the smallest lag grid and $41 \times 41 \times 21$ grid cells, we fitted the four variogram models to data from the 100 realisations of the four Gaussian random fields. The sill was estimated from data, and the exponent of the general exponential variogram was held fixed at 1.5 during all optimisations. Table 1 shows which variogram model that fitted data best in each case, in the sense that it had the smallest least-mean-squares residual.

The table shows that the Gaussian and the general exponential variograms are readily identified, including when only 10 wells are used. For the spherical variogram some 50 wells are needed for a positive identification, while the exponential variogram is falsely identified as spherical in 20 out of 100 cases when as many as 200 wells are used. As the exponential variogram gives less correlated fields than the spherical variogram, this may come as a result of smoothing of data. Note, however, that when 10 wells are used there is also a large number of fields having spherical variogram that are falsely identified as exponential.

Finally, it should be pointed out that if the exponents of the general exponential variogram had been allowed to vary freely during the optimisation, this variogram type would probably have given the best fit to most of the fields involved, with the possible exception of the fields having a spherical variogram.

5 Conclusions

Variogram estimation in 3D is surprisingly difficult. Even when 200 vertical wells are included in the parameter estimation, the total uncertainty involved is relatively large. However, based on our case study, we have come up with the following guidelines for minimising the uncertainty:

- The sill should be estimated directly from data rather than optimised.
- When empirical variogram estimates are made, the lag grid should be small; horisontal and vertical sizes equal to the respective ranges seem appropriate.
- The optimisation is not sensitive to lag grid binning as long as very coarse-meshed grids are avoided. Very fine-meshed grids should also be avoided to reduce computation time.

Table 2. The best RMSE measures obtained within the different data sets.

	1. range	2. range	3. range	azimuth	dip	sill	α
10 wells	660	180	2.5	35	6.3	0.26	0.12
50 wells	81	41	1.5	17	1.8	0.15	0.15
100 wells	41	34	1.5	10	1.0	0.14	0.18
200 wells	30	22	1.4	5	0.5	0.13	0.12
orig. value	200	100	10	60	3	1	1.5

The second of these guidelines is based on the fact that the (100m, 100m, 4m) grid generally gave better parameter estimates. The RMSEs obtained with this lag grid and the $41 \times 41 \times 21$ binning, are listed in Table 2. The values were obtained with the spherical and general exponential (α only) variograms.

When 50 wells were used, our estimator gave smaller RMSE measures than the zero estimator for all parameters; and for the vertical range, the estimator did better also with 10 wells. If we compare the RMSE measures with the original parameter values, however, we conclude that some 200 vertical wells are needed for obtaining reliable estimates of the azimuth and dip anisotropy angles, while some 50 wells are sufficient for the horizontal ranges and the sill. For the vertical range 10 wells are sufficient.

Acknowledgements

The authors would like to thank dr. Ragnar Hauge for helpful comments during the preparation of this paper.

References

Chen, Yongliang and Jiao, Xiguo, *Semivariogram fitting with linear programming*, Computers & Geosciences, vol. 27, No 1, 2001, p. 71–76.

Cressie, Noel A., *Fitting Variogram Models by Weighed Least Squares*, Mathematical Geology, vol. 17, no. 5, 1985, p. 563–586.

Gill, Philip E. and Murray, Walter and Wright, Margaret H., *Practical Optimization*, Academic Press, 1981.

Gringarten, Emmanuel and Deutsch, Clayton V., *Teacher's Aide Variogram Interpretation and Modeling*, Mathematical Geology, vol. 33, no. 4, 2001, p. 507–534.

Hawkins, D.M. and Cressie, Noel A., *Robust Kriging - A Proposal*, J. Int. Assoc. Math. Geol., vol. 16, no. 1, 1984, p. 3–18.

Matheron, G., , Econ. Geol., vol. 58, 1963, p. 1246–1266.

Pardo-Igúzquiza, Eulogio, *VARFIT: a Fortran-77 program for fitting variogram models by weighted least squares*, Computers & Geosciences, vol. 25, No 3, 1999, p. 251–261.

Webster, R. and Oliver, M. A, *How large a sample is needed to estimate the regional variogram adequately?*, *In Proceedings of GEOSTATISTICS TRÓIA '92*, vol. 1, p. 155–166, ed. A. Soares, Kluwer Academic Publishers, 1992.

Yaglom, A. M., *Correlation theory of stationary and related random functions*, Springer series in statistics, 1987.

COMPARISON OF STOCHASTIC SIMULATION ALGORITHMS IN MAPPING SPACES OF UNCERTAINTY OF NON-LINEAR TRANSFER FUNCTIONS

SUMAIRA EJAZ QURESHI AND ROUSSOS DIMITRAKOPOULOS
WH Bryan Mining and Geology Research Centre, University of Queensland, Brisbane, Australia

Geostatistical simulations are routinely used to quantify the uncertainty in forecasts (responses) generated from any non-linear function of spatially varying parameters. The ability to map the uncertainty in these responses is critical. A comparison of sequential Gaussian simulation (SGS), sequential indicator simulation (SIS) and probability field simulation (PFS) is made in this study, using an exhaustive dataset sampled with a random stratified grid and three transfer functions, namely, minimum cost network flow, threshold proportion and geometric mean. The results show that SGS and SIS have comparable performance in terms of bias and precision while PFS performs less well in most cases. Increased data leads usually to better precision but not necessarily bias. The performance of the simulation methods in mapping spaces of uncertainty depends on the complexity of the transfer function, and that is not necessarily a well-understood aspect of the modelling process.

1. Introduction

Spatial uncertainty and risk analysis in earth science and engineering applications, including operations in fields such as petroleum, mining and environment, can be assessed using stochastic simulation methods coupled with generally non-linear transfer functions, in the form of mathematical models. These models may be operations research algorithms for mine production scheduling, three-phase flow equations in petroleum reservoirs for forecasting production, or complex classification functions for the remediation of contaminated land. Understanding how the simulation algorithms interact with these non-linear mathematical models in mapping risk in output parameters of interest (termed response) is important and identified in pertinent studies. For example, studies include optimising mine designs and related net present value assessment based on geological uncertainty (e.g., Dimitrakopoulos et al., 2002), petroleum reservoir forecasting and production analysis (e.g., Walcott and Chopra, 1991), or minimising risk in contaminated site assessment and remediation (e.g., Qureshi, 2002), and others.

A series of responses generated from the application of a transfer function on a set of realizations from stochastic simulations may be expressed as a map/description of the space of uncertainty of the responses. This form of expression provides the means for optimal decision-making and risk management. It is important to note that, for non-linear transfer functions, (i) an average type spatial map of the input parameter(s) does

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 959–968.

not provide an average expected map of the space of response uncertainty, thus there is a substantial technical reason to use stochastic simulation rather than estimation; and (ii) methods used to stochastically simulate descriptions of pertinent attributes must be evaluated in terms of the map of the uncertainty of the response, rather than the maps of the description of the attributes. The latter point suggests that comparing commonly used simulation algorithms in combination with non-linear transfer functions is of interest, as recognised in the past (e.g., Gotway and Rutherford, 1994).

Some of the commonly used stochastic simulation methods are compared in this study. Specifically, the performance of sequential Gaussian simulation (SGS), sequential indicator simulation (SIS) and probability field simulation (PFS) (Goovaerts, 1997) is assessed in terms of mapping spaces of uncertainty for non-linear transfer functions. The functions considered are the minimum cost network flow, threshold proportion, and geometric mean. The performance of the combinations is assessed using two different sample sets (one of 250 and one of 500) generated from the exhaustive Walker Lake data set (Isaaks and Srivastava, 1989) using random stratified sampling. The results are then compared with those of the exhaustive data set. Bias and precision of response uncertainty distributions are the two quantities used for the comparison, which is based on the generation of 100 realizations for each sample dataset and method being compared. Note that, for the study presented here, the exhaustive data set and simulated realizations represent a 260m x 300m grid of 78,000 points, which are converted into 500 blocks (20x25 nodes per block) prior to input to the minimum cost path transfer functions.

In the following sections, the transfer functions and criteria for comparison are outlined; then the mapping of response uncertainty is presented. The results obtained are then compared; followed by a summary and conclusions.

2. Transfer functions, response uncertainty and criteria for comparison

Three transfer functions are used in this study: the mean of geometric means, threshold proportions, and minimum cost path. The functions are briefly described below, before the methods for the analysis and evaluation of results are described.

2.1. MEAN OF GEOMETRIC MEANS TRANSFER FUNCTION

For each interior grid node, the geometric mean (GM) is first computed as:

$$y_j = \exp\left[\frac{1}{m_j}\int_{m_j} \ln(g_{m_j})dg_{m_j}\right] \qquad j = 1,\ldots,N$$

where y_j = geometric mean at node j, m_j = 25 closest nodes to node j, g_{m_j} = 25 simulated values closest to node j, N = total number of nodes. For the measurements, which have zero or negative values, a constant is added to ensure that all values are positive. Then, the mean of geometric means (MGM) is obtained by averaging all the geometric means as:

$$y = \frac{1}{N}\int_N y_j dy_j$$

2.2. THRESHOLD PROPORTION TRANSFER FUNCTION

The threshold proportion (TP) transfer function is the proportion of the values above a specified threshold, i.e. the proportion of the values greater than the Pth percentile of the exhaustive data set. For this study, the 90th percentile is used as the threshold. If n = number of data greater than Pth percentile (i.e. 90th percentile) of the exhaustive data set and N = total number of data, then $\mathrm{TP} = \frac{n}{N}$

2.3. MINIMUM COST PATH TRANSFER FUNCTION

For the minimum cost path transfer function, the exhaustive data set and simulated realizations of 78,000 points are first converted into 500 blocks. In this path, a particle is released from the upper left corner and allowed to move horizontally from left to right, vertically downward or diagonally downward towards the lower right corner. Movement costs are based on the reciprocals of the block-averaged values. This transfer function is computed by using the minimum cost path network flow model (Qureshi, 2002). AMPL (Robert et al., 1993) is used to develop the required models for calculating the minimum cost paths.

2.4. MEASUREMENTS OF UNCERTAINTY DISTRIBUTIONS OF RESPONSE: BIAS AND PRECISION

Possibly the most common method for analysing the transfer functions and evaluating the results, is the bias measurement of response uncertainty distributions. Bias is measured as the absolute difference between the median of the uncertainty distribution and the true value, divided by the true value. Mathematically, if $\tilde{X}$ represents the median of the uncertainty distribution and $\bar{X}^{\mathrm{True}}$ represents the true value (obtained from the exhaustive data set), then the bias measurement can be written as:

$$\mathrm{Bias} = \frac{\left|\tilde{X} - \bar{X}^{\mathrm{True}}\right|}{\bar{X}^{\mathrm{True}}}$$

Lower numbers of the bias measure are indicative of uncertainty distributions whose values are consistently close to those computed from the exhaustive data set.

Precision may be seen as a measure of the magnitude of closeness of agreement among individual measurements. Precision is measured as the difference between the 90th percentile and 10th percentile of each uncertainty distribution divided by the corresponding percentile difference of the uncertainty distribution obtained by using SGS[*] with the 250 sample set. If X_{10} denotes the 10th percentile and X_{90} denotes the 90th percentile of uncertainty distributions, and if $X_{10}^{\mathrm{Strata\,250(SGS)}}$ and $X_{90}^{\mathrm{Strata\,250(SGS)}}$ denote the 10th and 90th percentiles respectively of the uncertainty distributions obtained by using SGS with the 250 sample set, then mathematically precision can be expressed as:

$$\mathrm{Precision} = \frac{X_{90} - X_{10}}{X_{90}^{\mathrm{Strata\,250(SGS)}} - X_{10}^{\mathrm{Strata\,250(SGS)}}}$$

[*] Percentile difference of uncertainty distributions can be divided by any percentile difference of uncertainty distributions obtained using any simulation algorithm with any sample set.

3. Mapping response uncertainty

The results of this study show a number of differences between the three simulation algorithms being compared. To help assess the methods, each response uncertainty distribution is compared to the true value computed from the Walker Lake exhaustive data set. For each of the simulation algorithms, response uncertainty distributions are compared with respect to sample size to see the effect of increasing the sample size on the resulting distributions.

3.1. UNCERTAINTY DISTRIBUTIONS BASED ON SGS

The SGS-produced uncertainty distributions obtained from the 250 and 500 sample sets are shown in Figure 1. In the figure, the horizontal line denotes the true transfer function value. Measures of the bias and precision of the response uncertainty distributions are summarised in Table 1.

Table 1. Summary of uncertainty distributions (produced by SGS) combined across all the transfer functions and sample sets

Sample Size	**Transfer Function**	**Median**	**True Value**	**Bias**	**X_{10}***	**X_{90}***	**Precision**	**Mean**	**SD***
250	G-Means*	263.46	247.50	0.06	251.26	274.97	1.00	263.67	8.62
	T-Proportion*	12.96	10.47	0.24	11.64	14.12	1.00	12.99	0.95
	Min-Cost*	6.66	6.96	0.04	5.92	7.81	1.00	6.94	1.21
500	G-Means	266.52	247.50	0.08	259.30	272.63	0.56	266.39	5.32
	T-Proportion	10.43	10.47	0.00	9.67	11.32	0.67	10.46	0.68
	Min-Cost	6.24	6.96	0.10	5.85	6.69	0.44	6.27	0.39

G-Means* = Mean of geometric mean transfer function, T-Proportion* = Threshold proportion transfer function, Min-Cost* = Minimum cost path transfer function, , X_{10}* = 10th percentile, X_{90}* = 90th percentile, SD* = Standard deviation

The uncertainty distribution obtained from the mean of geometric means transfer function based on the 250 sample set is less precise (i.e. 1.00) than the distribution obtained from the 500 sample set (see Figure 1 and Table 1). The uncertainty distribution obtained from threshold proportion transfer function using the 250 sample set is also less precise than the distribution based on the 500 sample set. For the 500 sample set, the threshold proportion transfer function is working really well because the uncertainty distribution obtained from this transfer function is unbiased. On the other hand, although the uncertainty distribution obtained from the minimum cost path using the 500 sample set is more precise than the distribution obtained from the 250 sample set, it has higher bias.

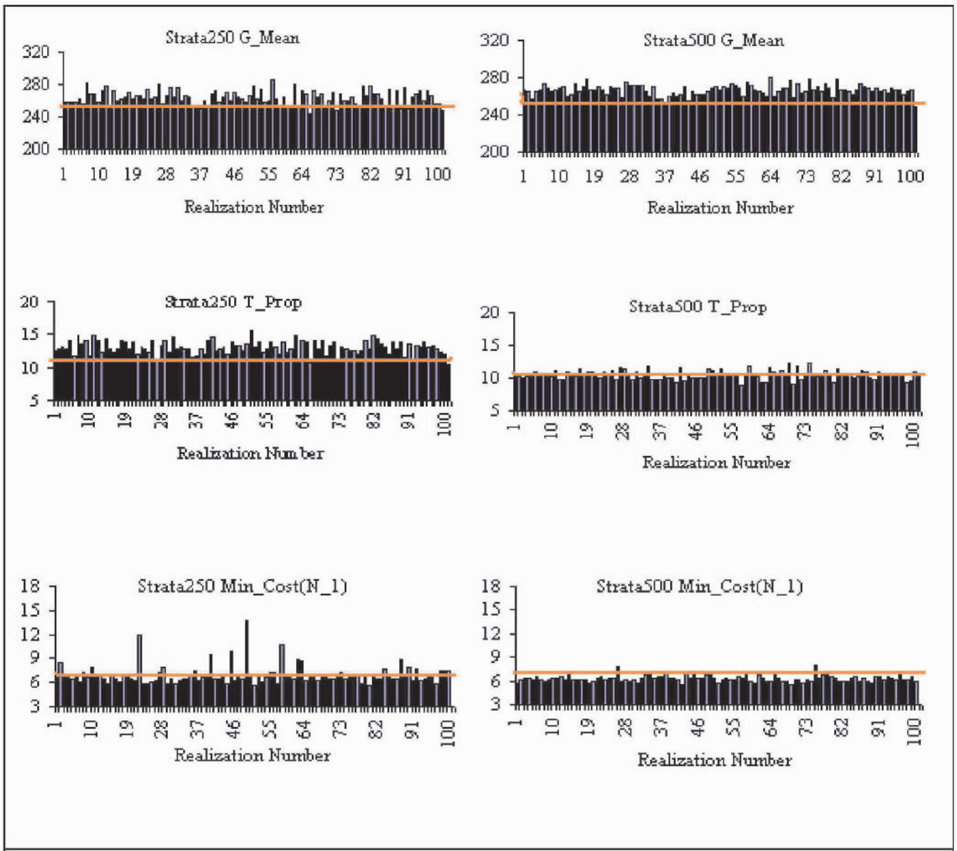


Figure 1. Uncertainty distributions for transfer functions obtained from realizations based on 250 and 500 sample sets, generated by SGS. The horizontal line represents the true transfer function value.

3.2. UNCERTAINTY DISTRIBUTIONS BASED ON SIS

The SIS produced uncertainty distributions based on the 250 and 500 sample sets are presented in Figure 2. Measures of the bias and precision of the response uncertainty distributions are summarised in Table 2.

Table 2. Summary of uncertainty distributions (produced by SIS) combined across all the transfer functions and sample sets

Sample Size	Transfer Function	Median	True Value	Bias	X_{10}*	X_{90}*	Precision	Mean	SD*
250	G-Means	248.57	247.50	0.00	237.21	258.97	0.92	248.19	8.04
	T-Proportion	12.69	10.47	0.21	11.30	14.08	1.12	12.75	1.03
	Min-Cost	6.83	6.96	0.02	5.90	7.52	0.86	6.74	0.63
500	G-Means	244.77	247.50	0.01	235.60	252.10	0.70	244.21	6.71
	T-Proportion	11.52	10.47	0.10	10.64	12.77	0.86	11.61	0.85
	Min-Cost	6.22	6.96	0.11	5.66	6.58	0.49	6.16	0.36

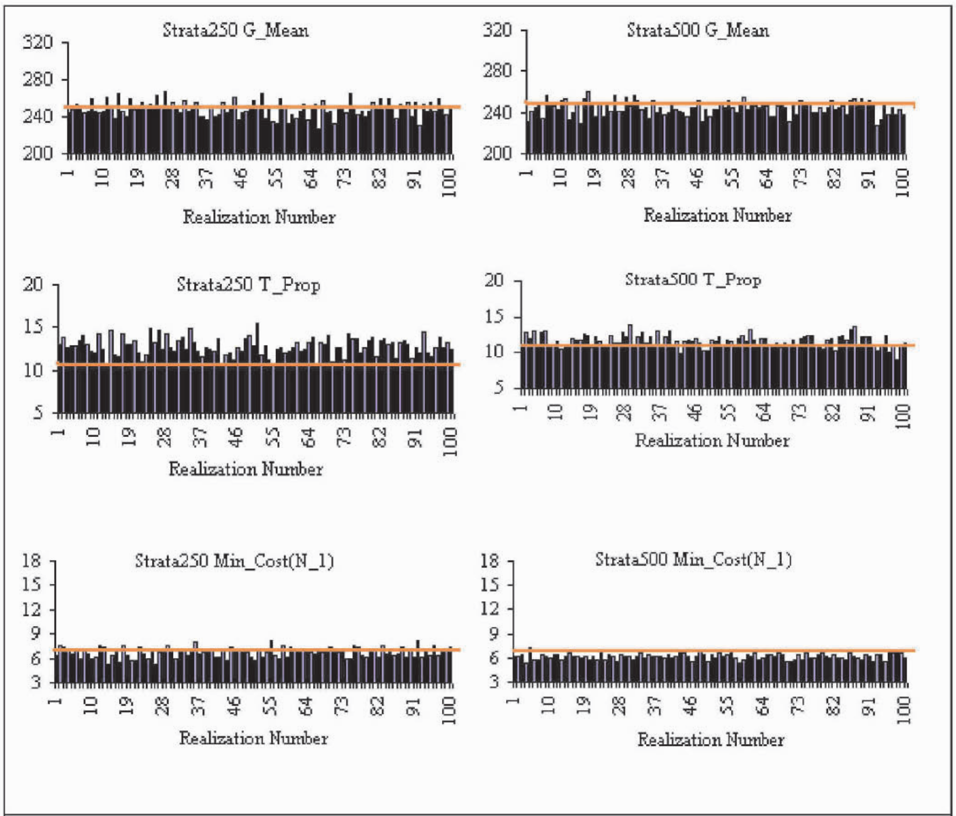


Figure 2. Uncertainty distributions for transfer functions obtained from realizations based on 250 and 500 sample sets, generated by SIS. The horizontal line represents the true transfer function value.

For the mean of geometric means and minimum cost path transfer functions, the uncertainty distribution obtained from the 250 sample set is less precise but has lower bias than the distribution produced using the 500 sample set. It would appear that increase in precision due to increase in sample size is obtained at the expense of bias.

For the threshold proportion transfer function, precision of the uncertainty distributions also increases with increase in sample size. However, for the same transfer function, the uncertainty distribution based on the 250 sample set has higher bias than the distribution obtained from 500 sample set.

Uncertainty distributions based on SGS and SIS using both the 250 and 500 sample sets are acceptable in terms of bias and precision. On average, for the uncertainty distributions obtained from both SGS and SIS, the increase in precision due to increase in sample size tends to increase bias in the resulting response distributions.

3.3. UNCERTAINTY DISTRIBUTIONS BASED ON PFS

The PFS-produced uncertainty distributions based on the two sample sets are presented in Figure 3. The measures of bias and precision of the response uncertainty distributions are summarised in Table 3.

Table 3. Summary of uncertainty distributions (produced by PFS) combined across all the transfer functions and sample sets

Sample Size	Transfer Function	Median	True Value	Bias	X_{10}*	X_{90}*	Precision	Mean	SD*
250	G-Means	272.55	247.50	0.10	228.98	312.46	3.52	273.39	31.77
	T-Proportion	13.00	10.47	0.24	9.00	17.31	3.35	13.36	3.04
	Min-Cost	7.11	6.96	0.02	5.77	10.20	2.34	7.67	1.82
500	G-Means	285.56	247.50	0.15	241.13	325.30	3.55	283.06	31.59
	T-Proportion	12.69	10.47	0.21	9.01	15.98	2.81	12.56	2.72
	Min-Cost	6.68	6.96	0.04	5.37	9.98	2.44	7.84	4.13

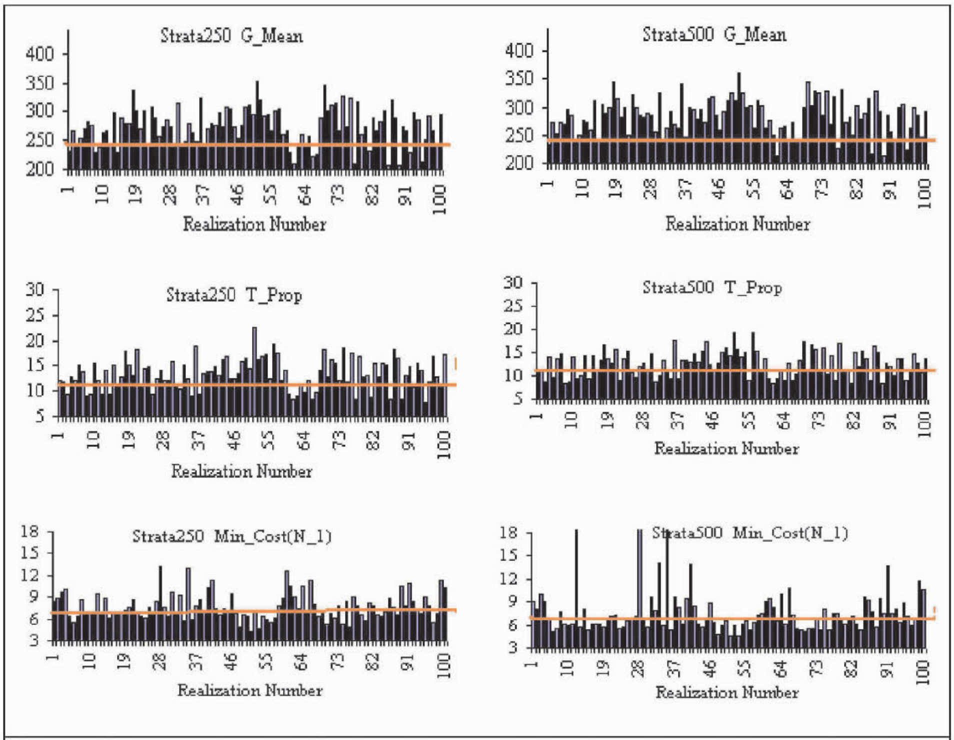


Figure 3. Uncertainty distributions for transfer functions obtained from realizations based on 250 and 500 sample sets, generated by PFS. The horizontal line represents the true transfer function value.

For the geometric means and minimum cost path transfer functions, the increase in the sample size results in a decrease in the precision of the uncertainty distribution. However, the uncertainty distributions based on the threshold proportion transfer function are reasonable for both sample sets, and much better than the distributions based on the other two transfer functions. For this transfer function, the precision increased and bias decreased as the sample size increased.
For the mean of geometric means transfer function, the uncertainty distribution based on the 500 sample set is low in precision but high in bias as shown in the Table 3. For the minimum cost path transfer function, PFS, unlike SGS and SIS, produces uncertainty distributions in which precision decreases with increase in sample size.

The uncertainty distributions produced by PFS are visually different from those produced by the sequential simulation algorithms, with the latter tending to exhibit more clustering of similar values. In most cases for PFS, the uncertainty distributions are less precise than the distributions obtained using SGS and SIS. Note that in most cases (Table 3) the average value of the response (Column 9) is higher than the true value (Column 4).

4. Comparative analysis

The measures of bias and precision of uncertainty distributions of response, based on both the 250 and 500 sample sets and using SGS, SIS and PFS, are combined in Figures 4 and 5 and 6. The reason for combining all these results is to see which algorithm overall is working well in the designed comparative study.

For the mean of geometric means transfer function (Figure 4), the uncertainty distributions based on SGS and SIS are highly precise and very similar to each other, compared with the distributions based on PFS. In all the cases, increase in sample size increases the precision associated with the uncertainty distributions, but there is no guarantee of improvement in bias measurements.

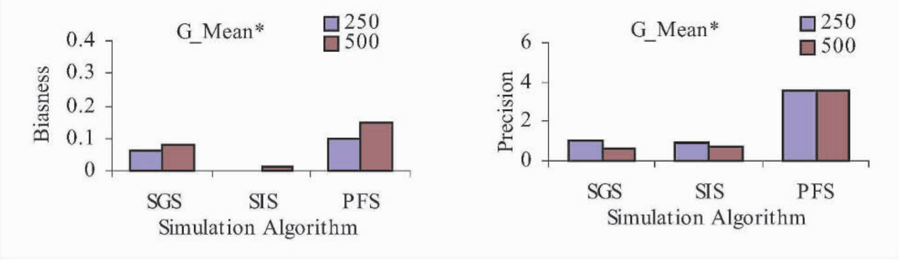


Figure 4. Bias and precision of uncertainty distributions obtained from mean of geometric means transfer functions using 250 and 500 sample sets.

The uncertainty distributions based on SGS and SIS using the threshold proportion transfer function (Figure 5) are more precise than the distributions based on PFS. For the minimum cost path transfer function (Figure 6), the uncertainty distributions based on SGS and SIS are more precise than the distributions based on PFS.

G_Mean* = Mean of geometric mean transfer function, T_Prop* = Threshold proportion transfer function

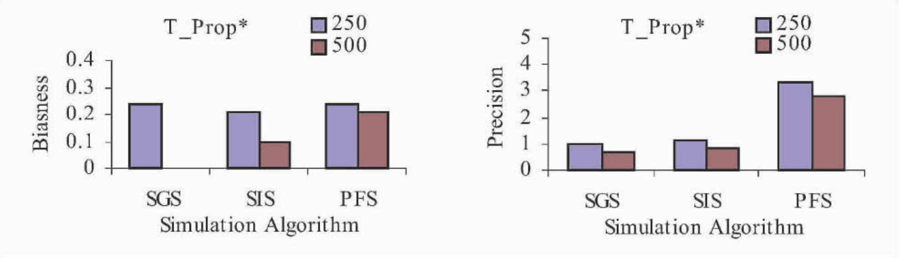


Figure 5. Bias and precision of uncertainty distributions obtained from threshold proportion transfer functions using 250 and 500 sample sets.

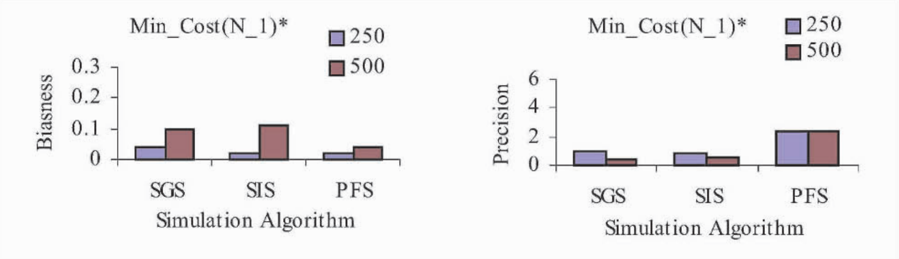


Figure 6. Bias and precision of uncertainty distributions obtained from minimum cost path transfer function using 250 and 500 sample sets.

5. Summary and conclusions

The performance of SGS, SIS and PFS in mapping spaces of uncertainty for non-linear transfer functions was compared for three transfer functions: minimum cost network, threshold proportion, and geometric mean. For the comparison undertaken in this study, stratified random sampling was used to choose two sample sets—one of 250 points and one of 500 points—from the exhaustive Walker Lake dataset. For each of the two sample sets, the three stochastic simulation algorithms being compared were used to generate 100 conditional realizations for each simulation method. For each realization, transfer functions were computed, each giving one value of the response. The results of the experiment provided several uncertainty distributions for each simulation algorithm. Each uncertainty distribution corresponded to one of the combinations of sample set and transfer function. The uncertainty distributions were then compared to the single value obtained from the exhaustive data set, using methods for the analysis and evaluation of results presented in this study.

Several broad issues are illustrated by the results of this study. The first is the effect of increasing the sample size on the resulting uncertainty distribution. It is found in almost all the cases that increasing the sample size improves the precision associated with response distribution.

· G_Mean* = Mean of geometric mean transfer function, T_Prop* = Threshold proportion transfer function
· Min_Cost(N_1)* = Minimum cost path transfer function

Second, the results indicate that, overall, sequential Gaussian and indicator based simulation models can incorporate the essential features of a spatially varying parameter.

From the results, it is clear that the uncertainty distributions produced by sequential based simulation algorithms are more precise, and their response values are closer to the true values, than the distributions produced by probability field simulation algorithm.

6. References

Dimitrakopoulos, R., Farrelly, C. and Godoy, M.C. (2002). Moving forward from traditional optimisation: Grade uncertainty and risk effects in open pit mine design. Transactions of the IMM, Section A Mining Industry, v. 111, pp. A82-A89.

Gotway, C. A. and Rutherford, B. M. (1994). *Stochastic simulation for imaging spatial uncertainty: Comparison and evaluation of available algorithms*: in Armstrong, M. and Dowd, P. A. (Eds), Geostatistical Simulations. Kluwer Academic Publisher, Dordrecht. 1-21.

Goovaerts, P. (1997). *Geostatistics for Natural Resources Evaluation.* Oxford University Press, New York.

Isaaks, E. H. and Srivastava, R. M. (1989). *An introduction to Applied Geostatistics*. Oxford University Press, New York.

Qureshi, S. E. (2002). *Comparative study of simulation algorithms in mapping spaces of uncertainty.* MPhil thesis, University of Queensland. Brisbane.

Robert, F., David, M. G. and Brian, W. K. (1993). *AMPL : A Modelling Language for Mathematical programming*. The Scientific Press Series, Ferncroft Village.

Walcott, D. S. and Chopra, A. K. (1991). Investigating infill drilling performance and reservoir continuity using geostatistics: in Linville, B. (Ed), *Proceedings Third International Reservoir Characterization Technical Conference.* PennWell Books, Oklahoma, 297-326.

INTEGRATING MULTIPLE-POINT STATISTICS INTO SEQUENTIAL SIMULATION ALGORITHMS

JULIÁN M. ORTIZ AND XAVIER EMERY
Department of Mining Engineering,University of Chile,
Av. Tupper 2069, Santiago, Chile, 837 0451

Abstract. Most conventional simulation techniques only account for two-point statistics via the modeling of the variogram of the regionalized variable or of its indicators. These techniques cannot control the reproduction of multiple-point statistics that may be critical for the performance of the models given the goal at hand (flow simulation in petroleum applications, planning and scheduling for mining applications).

Multiple-point simulation is a way to deal with this situation. It has been implemented for categorical variables, yet the demand of large data sets (training images) to infer the multiple-point statistics has impeded its use in the case of continuous variables.

We propose a method to incorporate multiple-point statistics into sequential simulation of continuous variables. Any sequential algorithm can be used. The method proceeds as follows. First, the multiple-point statistics are inferred from a training data set or training image with the typical indicator approach. The conditional probabilities given multiple-point data events enable to update the conditional distributions obtained by the sequential algorithm that uses the conventional two-point statistics. The key aspect is to preserve the shape of the conditional distribution between thresholds after updating the probability for the cutoffs used to infer the multiple-point statistics. Updating takes place under the assumption of conditional independence between the conditional probability obtained from the training set and the one retrieved from the conditional probability defined by the sequential method. The algorithm is presented in generality for any sequential algorithm and then illustrated on a real data set using the sequential indicator and Gaussian simulation methods. The advantages and drawbacks of this proposal are pointed out.

1 Introduction

Geostatistical simulation is being used increasingly for uncertainty quantification. Traditionally, simulation methods only rely on the inference and modeling of a variogram that characterizes the spatial continuity of the variable of interest (see for example Goovaerts, 1997; Chilès and Delfiner, 1999). However, in most real applications the variogram cannot capture some important features of the true variable. The main reason for the poor performance of models built using conventional simulation tools is that variogram models only control the joint behavior of pairs of

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 969–978.

points and there is no explicit control on the joint behavior of multiple points. The algorithm used and its underlying assumptions dictates how relationships between multiple points are controlled. Two-point statistics such as the variogram or covariance are not enough to describe some complex features that the real phenomenon may present.

This problem was addressed for categorical variables by Guardiano and Srivastava (1993) in the early nineties. They introduced the idea of going beyond bivariate moments, through the use of *extended normal equations*. The method was based on using a training image for inferring the multiple-point indicator frequencies and then drawing an indicator value for the categorical variable given the probability of the unknown node to belong to each category (facies). This approach was improved by the implementation of Strebelle and Journel (2000) called *single normal equation*, where a search tree was used to find the multiple-point frequencies. Deutsch (1992) proposed the integration of multiple-point statistics in a *simulated annealing* framework. Caers and Journel (1998) used *neural networks* to infer the conditional distributions in a non-linear fashion considering multiple-point statistics. Both authors relied on the use of training images and their applications were oriented to categorical variables.

More recently, Ortiz and Deutsch (2004) suggested the use of multiple-point statistics extracted from production data (blast hole data in mining applications). These statistics are integrated into sequential indicator simulation. The probability of an unsampled location to belong to a class defined by two cutoffs can be approximated using the probability obtained by conventional indicator kriging or using the probability estimated from the training data for the same configuration of class grades in nearby informed locations. These two statistics are then combined under the assumption that they are conditionally independent (Journel, 2002).

In this article, we extend the approach proposed by Ortiz and Deustch (2004) to integrate multiple-point statistics in any sequential simulation algorithm. The key aspect is to infer the conditional distribution and then update it only at few thresholds, preserving its shape as much as possible. The proposed approach is implemented on a case study, where two benches of a copper mine are simulated using sequential Gaussian and sequential indicator simulation and then updated using multiple-point statistics.

2 Inferring multiple-point statistics

Inference of multiple-point statistics is a difficult problem and requires having abundant data over a large domain. Furthermore, these should be regularly spaced to allow repetition of patterns of several points. In practice, this problem has been solved using training images. Alternatively, in mining applications, the use of abundant and pseudo-regular production data from samples taken in blast holes can replace the training image (Ortiz, 2003; Ortiz and Deutsch, 2004).

Inference of the multiple-point statistics is done using the indicator coding. First, a number of thresholds are defined and a multiple-point pattern is used to scan the training image or training data set. For each threshold, the training data are coded as one if they belong to the corresponding class, that is, if the value is lower than or equal to the corresponding threshold value, and zero otherwise. From the scanning of the training image or data set, the probability of a node being less than or equal to the threshold can be calculated based on the experimental frequencies of that event.

Since there is no modeling of the experimental multiple-point frequencies, an important limitation of this procedure is that the training data set and modeling scale must be equivalent. That is, the spacing of the (pseudo-)regular data in the training set must be identical to the spacing between nodes that are simulated subsequently. A second problem of this data-driven approach is that mathematical inconsistencies between statistics inferred from the training data and from the sample data used to condition the simulation may exist.

3 Updating conditional distributions with multiple-point statistics

The proposed approach to update conditional distributions with multiple-point statistics consists of the following steps:

1. Define a random path to visit the nodes in the simulation grid.
2. At every visited node, determine the conditional distribution by simple kriging of the (coded) sample data and previously simulated nodes.
3. Discretize the conditional distribution by a set of thresholds, which are interpreted as the conditional probability of the variable at that location not to exceed the corresponding threshold value.
4. Update the conditional probabilities originated from discretizing the conditional distribution by assuming conditional independence between them and the probability of a node to exceed the corresponding threshold given the multiple-point configuration of (coded) original sample data and previously simulated nodes.
5. Fill in the discretized conditional distribution using some interpolation rule and, more importantly, extrapolation of the tails.
6. Draw a uniform random value in [0,1] to read from the conditional distribution a simulated value.
7. Proceed to the next node in the random path until all nodes have been simulated.

The updating technique described in step 4 was presented by Journel (2002) under the name of *permanence of ratios assumption*, but it is equivalent to the well-known conditional independence assumption used in the Naïve Bayes classifiers (Warner et al, 1961; Anderson, 1974; Friedman, 1997).

This methodology can be applied to any sequential simulation algorithm where the conditional distribution at the simulation nodes has been defined. The most straightforward approach would be to apply it in an indicator context (Ortiz and Deutsch, 2004). In the following sections we present the details of implementing this method using indicator, Gaussian, isofactorial and direct simulation.

4 Implementation

Let the event **A** be the probability of a node not to exceed a threshold. Event **B** is defined by the indicator coded information provided by *n* single point events: $\{I(\mathbf{u}_1)=i_1, I(\mathbf{u}_2)=i_2,\ldots,I(\mathbf{u}_n)=i_n\}$. Finally, event **C** is the multiple-point event defined by the values of the indicators of *m* points: $\{I(\mathbf{u}'_1)=i'_1, I(\mathbf{u}'_2)=i'_2,\ldots,I(\mathbf{u}'_m)=i'_m\}$ (some of the *n* points belonging to **B** may also be part of **C**).

Indicator, Gaussian, isofactorial, or direct simulation can provide a conditional distribution that allows the calculation of *P*(**A** | **B**). The training dataset provides an estimate of *P*(**A** | **C**). Obtaining *P*(**A** | **B**, **C**) requires knowing the relationship between **B** and **C**, which is generally extremely difficult to get. Some assumption is required. These probabilities are combined assuming they are conditionally independent given **A**, that is, considering the expression for *P*(**A** | **B**, **C**) and $P(\overline{\mathbf{A}}\,|\,\mathbf{B},\mathbf{C}) = 1 - P(\mathbf{A}\,|\,\mathbf{B},\mathbf{C})$:

$$P(\mathbf{A}\,|\,\mathbf{B},\mathbf{C}) = \frac{P(\mathbf{A})\cdot P(\mathbf{B}\,|\,\mathbf{A})\cdot P(\mathbf{C}\,|\,\mathbf{A},\mathbf{B})}{P(\mathbf{B},\mathbf{C})} \qquad P(\overline{\mathbf{A}}\,|\,\mathbf{B},\mathbf{C}) = \frac{P(\overline{\mathbf{A}})\cdot P(\mathbf{B}\,|\,\overline{\mathbf{A}})\cdot P(\mathbf{C}\,|\,\overline{\mathbf{A}},\mathbf{B})}{P(\mathbf{B},\mathbf{C})}$$

and the conditional independence conditions

$$P(\mathbf{B}\,|\,\mathbf{A},\mathbf{C}) = P(\mathbf{B}\,|\,\mathbf{A}) \text{ and } P(\mathbf{C}\,|\,\mathbf{A},\mathbf{B}) = P(\mathbf{C}\,|\,\mathbf{A})$$

the ratio between *P*(**A** | **B**, **C**) and its complement $P(\overline{\mathbf{A}}\,|\,\mathbf{B},\mathbf{C})$ can be calculated as:

$$\frac{P(\mathbf{A}\,|\,\mathbf{B},\mathbf{C})}{P(\overline{\mathbf{A}}\,|\,\mathbf{B},\mathbf{C})} = \frac{P(\mathbf{A})\cdot P(\mathbf{B}\,|\,\mathbf{A})\cdot P(\mathbf{C}\,|\,\mathbf{A},\mathbf{B})}{P(\overline{\mathbf{A}})\cdot P(\mathbf{B}\,|\,\overline{\mathbf{A}})\cdot P(\mathbf{C}\,|\,\overline{\mathbf{A}},\mathbf{B})} = \frac{P(\mathbf{A})\cdot P(\mathbf{B}\,|\,\mathbf{A})\cdot P(\mathbf{C}\,|\,\mathbf{A})}{P(\overline{\mathbf{A}})\cdot P(\mathbf{B}\,|\,\overline{\mathbf{A}})\cdot P(\mathbf{C}\,|\,\overline{\mathbf{A}})}$$

This expression can be simplified to:

$$P(\mathbf{A}\,|\,\mathbf{B},\mathbf{C}) = \frac{\dfrac{1-P(\mathbf{A})}{P(\mathbf{A})}}{\dfrac{1-P(\mathbf{A})}{P(\mathbf{A})} + \dfrac{1-P(\mathbf{A}\,|\,\mathbf{B})}{P(\mathbf{A}\,|\,\mathbf{B})}\cdot\dfrac{1-P(\mathbf{A}\,|\,\mathbf{C})}{P(\mathbf{A}\,|\,\mathbf{C})}}$$

It can be seen that, under the assumption of conditional independence, the probability of event **A** can be calculated with relative ease, since it does not require knowing the relationship between **B** and **C**. We now present four cases where this approximation can be implemented.

4.1 INDICATOR SIMULATION

Consider the usual indicator coding:

$$i(\mathbf{u}_\alpha; z_k) = \begin{cases} 1, \text{ if } z(\mathbf{u}_\alpha) \le z_k \\ 0, \text{ otherwise} \end{cases} \qquad k = 1,\ldots,K$$

where $z(\mathbf{u}_\alpha)$ is the value at location $\mathbf{u}_\alpha$. This can be interpreted as a probability:

$$P(\mathbf{A}) = i(\mathbf{u}_\alpha; z_k) = \text{Prob}\{z(\mathbf{u}_\alpha) \le z_k\} = F(z_k)$$

For a simulated node located at $\mathbf{u}_0$, the conditional probability given the data in a search neighborhood can be calculated by simple indicator kriging (Journel, 1983; Alabert, 1987):

$$\begin{aligned} P(\mathbf{A} \mid \mathbf{B}) = \left[i(\mathbf{u}_0; z_k)\right]^*_{SIK} &= \left[\text{Prob}\{z(\mathbf{u}_0) \le z_k \mid (n)\}\right]^*_{SIK} \\ &= \sum_{\alpha=1}^{n} \lambda_\alpha^{SIK}(\mathbf{u}_0; z_k) \cdot i(\mathbf{u}_\alpha; z_k) + \left[1 - \sum_{\alpha=1}^{n} \lambda_\alpha^{SIK}(\mathbf{u}_0; z_k)\right] \cdot F(z_k) \end{aligned}$$

where (n) represents the conditioning information provided by n samples and previously simulated nodes in the search neighborhood, $\lambda_\alpha^{SIK}(\mathbf{u}_0; z_k)$ for $\alpha = 1,\ldots,n$ are the simple indicator kriging weights, and $F(z_k)$ is the global proportion below threshold z_k.

From the training information, a different conditional probability can be obtained for $z(\mathbf{u}_0)$ to be less than or equal to z_k, given the information of an m-point configuration:

$$P(\mathbf{A} \mid \mathbf{C}) = \left[\text{Prob}\{z(\mathbf{u}_0) \le z_k \mid (m)\}\right]^*_{MP}$$

The conditional probabilities obtained by simple indicator kriging can be updated with the conditional probabilities obtained from the training dataset, allowing the calculation of the conditional distribution that accounts for both, the set of n single point events and the single m-points event. Since the discretization of the conditional distributions by the indicator approach is generally coarse, the updated conditional distribution will also be a coarse approximation of the conditional distribution. The usual interpolation between the estimated indicators and extrapolation beyond the first and last thresholds is necessary (Deutsch and Journel, 1998).

4.2 GAUSSIAN SIMULATION

A natural extension to the implementation presented above is to update conditional probabilities obtained via a multigaussian sequential simulation. The method requires the transformation of the original distribution into a standard Gaussian distribution:

$$Z(\mathbf{u}) = \phi(Y(\mathbf{u}))$$

It is widely known that under the assumption of multivariate Gaussianity, the conditional distributions are fully defined by the mean and variance obtained by simple kriging, which requires the knowledge of the covariance function $C(\mathbf{h})$:

$$m_Y(\mathbf{u}_0) = \sum_{\alpha=1}^{n} \lambda_\alpha^{SK}(\mathbf{u}_0) \cdot y(\mathbf{u}_\alpha) \qquad \sigma_Y^2(\mathbf{u}_0) = 1 - \sum_{\alpha=1}^{n} \lambda_\alpha^{SK}(\mathbf{u}_0) \cdot C(\mathbf{u}_0, \mathbf{u}_\alpha)$$

The expression for the conditional distribution is:

$$P(\mathbf{A} \mid \mathbf{B}) = \left[\text{Prob}\{z(\mathbf{u}_0) \leq z_k \mid (n)\}\right]^*_{SK} = \left[\text{Prob}\{y(\mathbf{u}_0) \leq y_k \mid (n)\}\right]^*_{SK} \sim N\left(m_Y(\mathbf{u}_0), \sigma^2_Y(\mathbf{u}_0)\right)$$

An indicator-based approach to update these conditional distributions would consist in discretizing them into a series of thresholds, which can be easily done numerically, and then updating these conditional probabilities with the conditional probabilities obtained from the training set. Again, a decision about how to interpolate between the discrete points and beyond the first and last thresholds is necessary; however, in this case, the thresholds can discretize the conditional distribution more precisely than in indicator simulation. For instance, instead of choosing ten to fifteen thresholds, over a hundred thresholds can be easily taken, provided sufficient training information is available for reliable estimation of the conditional probabilities $P(\mathbf{A} \mid \mathbf{C})$.

4.3 ISOFACTORIAL SIMULATION

The next case of interest corresponds to sequential isofactorial simulation (Emery, 2002). Again, this method relies on a transformation of the original distribution into a new variable that follows a bivariate isofactorial distribution with marginal pdf $f(.)$. Notice that the transformation function may differ from the one used in the case of sequential Gaussian simulation and that $f(.)$ is not necessarily the standard Gaussian pdf. Typical applications consider transforming the raw variable to a Gaussian or Gamma distributions, although other cases can be considered, such as a Beta, Poisson, Binomial or Negative Binomial distribution. The conditional probability can be obtained by disjunctive kriging of the indicator function at a given threshold z_k:

$$P(\mathbf{A} \mid \mathbf{B}) = \left[i(\mathbf{u}_0; z_k)\right]^*_{DK} = \left[i(\mathbf{u}_0; y_k)\right]^*_{DK} = \sum_{p=0}^{\infty} \psi_p \cdot \left[\chi_p\left(Y(\mathbf{u}_0)\right)\right]^*_{SK}$$

where the coefficients are calculated as:

$$\psi_p = \int_{-\infty}^{y_k} \chi_p(y) \cdot f(y) \cdot dy$$

and each factor $\left\{\chi_p\left(Y(\mathbf{u}_0)\right), p \in \mathrm{N}\right\}$ is estimated by simple kriging from its values at the neighboring data locations:

$$\left[\chi_p\left(Y(\mathbf{u}_0)\right)\right]^*_{SK} = \sum_{\alpha=1}^{n} \lambda^{SK}_{\alpha,p}(\mathbf{u}_0) \cdot \chi_p\left(Y(\mathbf{u}_\alpha)\right)$$

The weights $\lambda^{SK}_{\alpha,p}(\mathbf{u}_0)$ are obtained by solving a simple kriging system considering a covariance function that depends on the isofactorial distribution and on the degree p. In practice only the first few factors are required (Matheron, 1976; Rivoirard, 1994; Chilès and Delfiner, 1999).

As with the conditional probabilities estimated by indicator kriging or under the multigaussian assumption, a probability conditional to the multiple-point event for each threshold can be estimated using the training data set, and subsequently used to update the conditional probability estimated under the isofactorial framework.

4.4 DIRECT SIMULATION

One last algorithm that could be considered is direct sequential simulation, which basically works by estimating the mean and variance of the conditional distribution by simple kriging. The shape of this distribution is then determined either by sampling the global distribution to match the mean and variance of the local conditional distribution (Soares, 2001), or by defining a conditional distribution lookup table (Oz et al., 2003). The procedure is virtually the same as in Gaussian simulation: obtain the conditional probability from the local distribution and update it with the multiple-point probability inferred from the training dataset.

5 Case study

The following case study presents some preliminary results about the application of the proposed methodology to simulate the point-support grades on a copper deposit, based on drill hole (exploration) information. The multiple-point statistics are extracted from production (blast hole) data obtained from two benches already mined out. This information is used to simulate the copper grade on two lower benches. An assumption of strict stationarity is required in order to "export" these multiple-point statistics. The example shows the updating technique implemented for the sequential indicator and Gaussian simulation algorithms.

Figure 1 shows the exploration data for a specific bench and the training information from one of the two benches used for multiple-point statistics inference. These statistics are inferred using a 5 points pattern made of a central node and the four adjacent nodes in the horizontal plane (no vertical data has been used for the multiple-point statistics inference). *Figure 2* displays realizations for a specific bench using sequential indicator and Gaussian simulation and the proposed methods where the conditional distributions are updated with multiple-point statistics extracted from the production data. The typical "patchiness" of indicator simulation appears clearly in the maps. This characteristic appears more strongly when multiple-point information is incorporated under the assumption of conditional independence. The patchiness disappears when using the Gaussian algorithm as a base method for inferring the conditional distributions: in this case, transitions from high to low grade zones are smoother. However, the integration of multiple-point statistics injects more connectivity to the realization.

Table 1 shows the total copper content and quantity of metal above a cutoff of 0.7 %Cu calculated over a particular area using the four methods for 20 × 20 × 12 m^3 panels. Smoother transitions and the added connectivity explain the higher variance obtained first, between indicator and Gaussian methods, and second, between the cases without and with multiple-point information. Validation remains a difficult issue and further research is required in this respect.

The implementation of these algorithms has shown some of the possible problems of their application. Numerical approximations are required to interpolate and extrapolate the tails once the discretization in indicators is performed for the updating procedure. Furthermore, the number of thresholds used depends on the quality and size of the training data set, in order to ensure reliable estimation of the multiple-point statistics.

	Cutoff = 0 %Cu		Cutoff = 0.7 %Cu	
	Mean	Std. Dev.	Mean	Std. Dev.
SISIM	69.25	1.43	67.12	1.71
SISIM-MP	70.88	1.72	67.44	2.06
SGSIM	69.85	1.64	67.54	2.00
SGSIM-MP	74.27	1.95	71.30	2.32

Table 1. Total quantity of metal above cutoffs of 0 and 0.7 %Cu from the sets of 100 realizations obtained with each method (in thousands of copper tonnes).

6 Conclusions

Integrating multiple-point statistics into sequential simulation algorithms can be achieved under some assumption of the relationship (redundancy) between the conditional probability inferred from a training data set, given a multiple-point event, and the conditional probability inferred by a conventional kriging approach (indicator, multigaussian, or disjunctive kriging). We propose assuming conditional independence between these two sources of information, to obtain an estimate of the conditional probability that accounts for the neighboring data (*n* points) and the closest multiple point configuration (*m* points). The updating methodology proposed can be applied to any sequential simulation algorithm, provided that a conditional distribution is calculated at each simulation node on the grid.

Implementation of this technique has proven challenging, particularly because of possible inconsistencies between the sources of information (biases) where the statistics are inferred, and because of numerical approximations (particularly when extrapolating the tails) required to obtain the simulated values from the updated conditional distributions. Furthermore, the assumption itself should be investigated. A model that accounts for the redundancy between the sources of information could be easily constructed by defining a parameter τ, such that: $P(\mathbf{C} \mid \mathbf{A})^{\tau} \approx P(\mathbf{C} \mid \mathbf{A}, \mathbf{B})$, hence:

$$\frac{P(\mathbf{A}\mid\mathbf{B},\mathbf{C})}{P(\overline{\mathbf{A}}\mid\mathbf{B},\mathbf{C})}\Big/\frac{P(\mathbf{A}\mid\mathbf{B})}{P(\overline{\mathbf{A}}\mid\mathbf{B})}=\left(\frac{P(\mathbf{A}\mid\mathbf{C})}{P(\overline{\mathbf{A}}\mid\mathbf{C})}\Big/\frac{P(\mathbf{A})}{P(\overline{\mathbf{A}})}\right)^{\tau}$$

However, the parameter τ is difficult to get and, to make things worse, it is location and data dependent.

The proposal in this article opens an interesting and original research avenue about the use of multiple-point statistics in a data-driven mode. Implementation and applications to real data will offer challenges that have yet to be discovered.

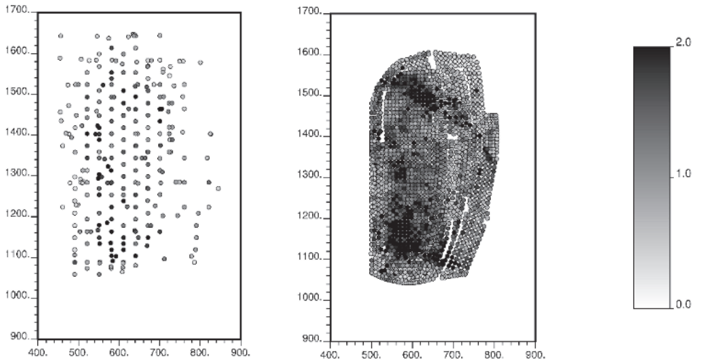


Figure 1. Left: Exploration (drill holes) data. Right: production (blast holes) data. Only the data in one bench are displayed. Production data are used to infer the multiple-point statistics.

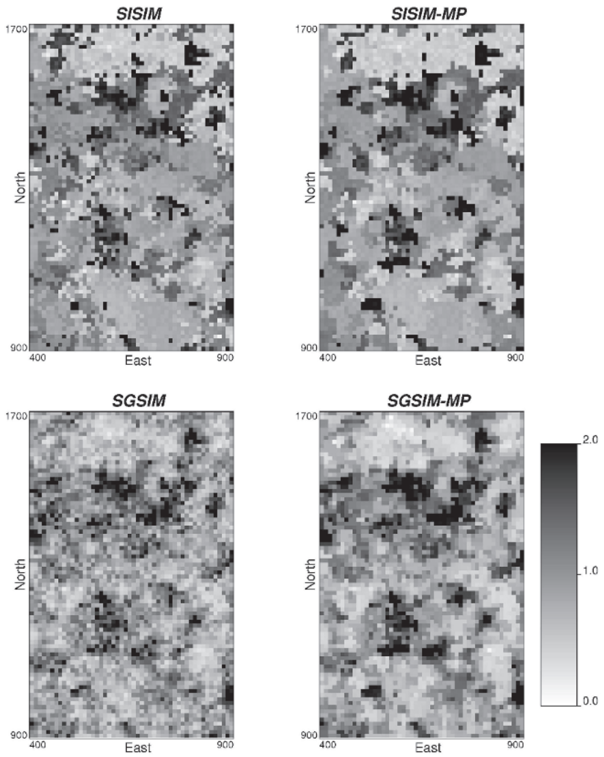


Figure 2. Plan views of a bench simulated using sequential indicator simulation updated with multiple-point statistics.

Acknowledgements

We would like to acknowledge the Department of Mining Engineering at the University of Chile and particularly its industry sponsor Codelco Chile for supporting this research.

References

Alabert, F. G., *Stochastic Imaging of Spatial Distributions using Hard and Soft Information*, MSc Thesis, Stanford University, Stanford, CA, 1987.

Anderson, J. A., Diagnosis by logistic discriminant function: further practical problems and results. *Applied Statistics*, vol. 23, no. 3, 1974, p. 397-404.

Caers, J. and Journel, A. G., Stochastic reservoir simulation using neural networks trained on outcrop data, *in* 1998 SPE Annual Technical Conference and Exhibition, New Orleans, LA. Society of Petroleum Engineers. SPE paper # 49026, 1998, p. 321-336.

Chilès, J. P. and Delfiner, P., *Geostatistics: Modeling spatial uncertainty*, Wiley, New York, 1999.

Deutsch, C. V., *Annealing Techniques Applied to Reservoir Modeling and the Integration of Geological and Engineering (Well Test) Data*. PhD thesis, Stanford University, Stanford, CA, 1992.

Deutsch, C.V. and Journel, A.G., *GSLIB: Geostatistical Software Library and User's Guide*, 2nd Edition, Oxford University Press, 1998.

Emery, X., Conditional simulation of non-Gaussian random functions. *Mathematical Geology*, vol. 34, no. 1, 2002, p. 79-100.

Friedman, J. H., On bias, variance, 0/1-loss, and the curse of dimensionality. *Data mining and knowledge Discovery*, vol. 1, 1997, p. 55-77.

Goovaerts, P., *Geostatistics for Natural Resources Evaluation*, Oxford University Press, 1997.

Guardiano, F. and Srivastava, M., Multivariate geostatistics: Beyond bivariate moments, *in* A. Soares, editor, Geostatistics Tróia'92. Kluwer, Dordrecht, vol. 1, 1993, p. 133-144.

Journel, A. G., Nonparametric estimation of spatial distributions, *Mathematical Geology*, vol. 15, no. 3, 1983, p. 445-468.

Journel, A. G., Combining knowledge from diverse sources: An alternative to traditional data independence hypotheses. *Mathematical Geology*, vol. 34, no. 5, 2002, p. 573-596.

Matheron, G., A Simple Substitute for Conditional Expectation: the Disjunctive Kriging, *in* M. Guarascio, M. David and C. Huijbregts, editors, Advanced Geostatistics in the Mining Industry. Reidel, Dordrecht, 1976, p. 221-236.

Ortiz, J. M., *Characterization of High Order Correlation for Enhanced Indicator Simulation*. PhD thesis, University of Alberta, Edmonton, AB, Canada, 2003.

Ortiz, J. M. and Deutsch, C. V., Indicator simulation accounting for multiple-point statistics. *Mathematical Geology*, vol. 36, no. 6, 2004, p. 545-565.

Oz, B., Deutsch, C. V., Tran, T. T. and Xie, Y., DSSIM-HR: A Fortran 90 program for direct sequential simulation with histogram reproduction. *Computers & Geosciences*, vol. 29, no. 1, 2003, p. 39-52.

Rivoirard, J., *Introduction to Disjunctive Kriging and Non-Linear Geostatistics*, Oxford University Press, New York, 1994.

Soares, A., Direct sequential simulation and cosimulation. *Mathematical Geology*, vol. 33, no. 8, 2001, p. 911-926.

Strebelle, S. and Journel, A. G., Sequential simulation drawing structures from training images, *in* 6th International Geostatistics Congress, Cape Town, South Africa, Geostatistical Association of Southern Africa, 2000.

Warner, H., Toronto, H., Veesey, L. and Stephenson, R., A mathematical approach to medical diagnosis. *Journal of the American Medial Association*, vol. 177, 1961, p. 177-183.

POST-PROCESSING OF MULTIPLE-POINT GEOSTATISTICAL MODELS TO IMPROVE REPRODUCTION OF TRAINING PATTERNS

SEBASTIEN STREBELLE[1] and NICOLAS REMY[2]
[1] *ChevronTexaco Energy Technology Company,*
6001 Bollinger Canyon Road, San Ramon, CA 94583, USA
[2] *Department of Geological and Environmental Sciences*
Stanford University, Stanford, CA 94305, USA

Abstract. In most petroleum and groundwater studies, flow performance is highly dependent on the spatial distributions of porosity and permeability. Because both porosity and permeability distributions primarily derive from facies deposition, facies should be the first property to be modeled when characterizing a reservoir. Yet, traditional geostatistical techniques, based on variogram reproduction, typically fail to model geologically-realistic depositional facies. Indeed, variograms only measure facies continuity between any two points in space; they cannot account for curvilinear and/or large-scale continuous structures, such as sinuous channels, that would require inferring facies joint-correlation at many more than two locations at a time.

Multiple-point geostatistics is a new emerging approach wherein multiple-point facies joint-correlation is inferred from three-dimensional training images. The simulation is pixel-based, and proceeds sequentially: each node of the simulation grid is visited only once along a random path, and simulated values become conditioning data for nodes visited later in the sequence. At each unsampled node, the probability of occurrence of each facies is estimated using the multiple-point statistics extracted from the training image. This process allows reproducing patterns of the training image, while honoring all conditioning sample data.

However, because of the limited size of the training image, only a very limited amount of multiple-point statistics can be actually inferred from the training image. Therefore, in practice, only a very few conditioning data close to the node to be simulated are used whereas farther away data carrying important large-scale information are generally ignored. Such approximation leads to inaccurate facies probability estimates, which may create "anomalies", for example channel disconnections, in the simulated realizations. In this paper, a method is proposed to use more data for conditioning, especially data located farther away from the node to be simulated. A measure of consistency between simulated realizations and training image is then defined, based on the number of times each simulated value, although initially identified as a conditioning datum to simulate a nearby node, had to be ignored eventually to be able to infer from the training image the conditional probability distribution at that node. Re-simulating the most inconsistent node values according to that measure enables improvement in the reproduction of training patterns without any significant increase of computation time. As an application, that post-processing process is used to remove channel disconnections from a fluvial reservoir simulated model.

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 979–988.

1 Introduction

Most geological environments are characterized by successive depositions of elements, or rock bodies, through time. These elements are traditionally grouped into classes, commonly named “depositional facies”, that correspond to particular combinations of lithology, physical and biological structures. For example, the typical depositional facies encountered in fluvial environments are high permeability sand channels, and sometimes, levies and splays with variable ranges of permeability.

Reservoir heterogeneity, hence flow performance, is primarily controlled by the spatial distribution of those depositional facies. Thus, a best practice would consist of modeling first depositional facies, and then populating each simulated facies with its corresponding specific porosity and permeability distributions. Yet traditional facies modeling techniques show severe limitations:

1. Variogram-based techniques, for example sequential indicator simulation (Deutsch and Journel, 1998), do not allow modeling geologically-realistic depositional elements because identification of two-point statistics, as modeled by the variogram, is not sufficient to characterize curvilinear or long-range continuous facies such as sand channels (Strebelle, 2000).
2. Object-based modeling techniques (Viseur, 1997) allow modeling quite realistic elements, but their conditioning is still commonly limited to a few wells.

An alternative technique proposed by Guardiano and Srivastava (1993) consists of going beyond the two-point statistics variogram by extracting multiple-point statistics from a training image. The training image can be defined as a three-dimensional conceptual geological model that depicts the geometry of each depositional facies expected to be present in the subsurface, as well as the complex spatial relationships existing among the different facies. Training images are typically obtained by interpreting available field data (cores, well logs, seismic), but also by using information from nearby field analogues and outcrop data. Figure 1a displays an example of a training image for a 2D horizontal section of a fluvial reservoir. That particular image was hand-drawn by a geologist, then numerically digitized. In 3D applications, three-dimensional training images are preferentially created using unconditional object-based modeling techniques.

Multiple-point statistics (MPS) simulation consists of reproducing patterns displayed in the training image, and anchoring them to the data actually sampled from the reservoir under study. In more detail, let S be the categorical variable (depositional facies) to be simulated, and s_k, k=1…K, the K different states (facies types) that the variable S can take. MPS simulation is a pixel-based technique that proceeds sequentially: all simulation grid nodes are visited only once along a random path and simulated node values become conditioning data for cells visited later in the sequence. At each unsampled node $\mathbf{u}$, let d_n be the data event consisting of the n closest conditioning data $S(\mathbf{u_1})=s(\mathbf{u_1})\ldots S(\mathbf{u_n})=s(\mathbf{u_n})$, which may be original sample data or previously simulated node values. The probability that the node $\mathbf{u}$ be in state s_k given d_n is estimated using Bayes’ relation:

$$\text{Prob}\{S(\mathbf{u}) = s_k \mid d_n\} = \frac{\text{Prob}\{S(\mathbf{u}) = s_k \text{ and } d_n\}}{\text{Prob}\{d_n\}}$$

Prob$\{S(\mathbf{u})=s_k$ and $d_n\}$ and Prob$\{d_n\}$ are multiple-point statistics moments that can be inferred from the training image:

1. Prob$\{d_n\}$=c(d_n)/N_{TI}, where N_{TI} is the size of the training image, and c(d_n) is the number of replicates of the conditioning data event d_n that can be found in the training image. By replicates, we mean training data events that have the same geometrical configuration and the same data values as d_n.
2. Prob$\{S(\mathbf{u})=s_k$ and $d_n\}$=c_k(d_n)/N_{TI}, where c_k(d_n) is the number of training replicates, among the c(d_n) previous ones, associated to a central value $S(\mathbf{u})$ in state s_k.

The conditional probability of occurrence of state s_k at location **u** is then identified as the proportion of state s_k obtained from the central values of the training d_n -replicates:

Prob$\{S(\mathbf{u})=s_k \mid d_n\}$=$\text{c}_\text{k}$($d_n$)/c($d_n$) (1)

The original MPS simulation implementation proposed by Guardiano and Srivastava was extremely cpu-time demanding since, at each node **u** to be simulated, the whole training image had to be scanned anew to search for training replicates of the local conditioning data event. Strebelle (2000) proposed decreasing the cpu-time by storing ahead of time all conditional probability distributions that can be actually inferred from the training image in a dynamic data structure called a search tree. More precisely, given a conditioning data search window *W*, which may be a search ellipsoid defined using GSLIB conventions (Deutsch and Journel, 1998), τ_N denotes the data template (geometric configuration) consisting of the *N* vectors $\{\mathbf{h}_\alpha, \alpha=1\ldots N\}$ corresponding to the *N* relative grid node locations included within *W*. Prior to the simulation, the training image is scanned with τ_N, and the numbers of occurrences of all the training data events associated with τ_N are stored in the search tree. During the simulation, at each unsampled node **u**, τ_N is used to identify the conditioning data located in the search neighborhood *W* centered on **u**. d_n denoting the data event consisting of the *n* conditioning data found in *W* (original sample data or previously simulated values, $n \leq N$), the local probability distribution conditioned to d_n is retrieved directly from the above search tree; the training image need not be scanned anew. Furthermore, to decrease the memory used to build the search tree and the cpu-time needed to retrieve conditional probabilities from it, a multiple-grid approach was implemented that consists of simulating a series of nested and increasingly-finer grids, and rescaling the data template τ_N proportionally to the node spacing within the grid being simulated (Tran, 1994; Strebelle, 2000). That multiple-grid approach enables the reproduction of the large-scale structures of the training image while keeping the size of the data template τ_N reasonably small ($N \leq 100$).

The MPS simulation program **snesim** (Strebelle, 2000) is applied to the modeling of a 2D horizontal section of a fluvial reservoir. The training image depicts the prior conceptual geometry of the sinuous sand channels expected to be present in the subsurface (Figure 1a). The size of that image is 250*250=62,500 pixels, and the channel proportion is 27.7%. An isotropic 40-data template was used to build the search trees for each of the four nested grids considered in the multiple-grid simulation approach. The unconditional MPS model generated by **snesim** reproduces reasonably well the patterns displayed by the training image, although some channel disconnections can be observed (Figure 1b).

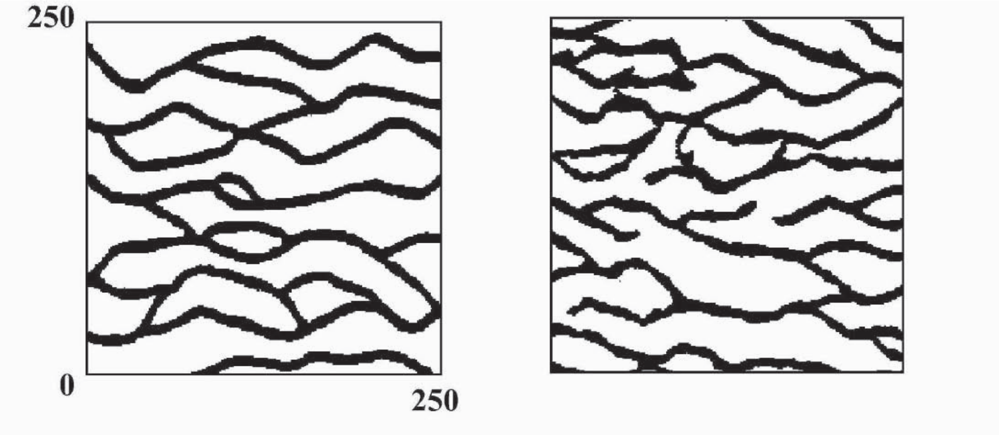


Figure 1. a) Training image used for the simulation of a 2D horizontal section of a fluvial reservoir (left); b) Corresponding MPS model generated by **snesim** (right)

If channel disconnections were believed to occur in the reservoir being modeled, due for example to some faults, the training image should display such disconnections. In the case above, the channel disconnections are not consistent with the information carried by the training image, thus they should be treated as "anomalies" that need to be corrected. In this paper we analyze why "anomalies" appear in MPS models. Then we propose modifying the **snesim** algorithm to decrease the number of anomalies, and we introduce a post-processing technique to remove those remaining.

2 Multiple-point statistics inference limitation

Inferring from the training image the probability conditional to a data event d_n requires finding at least one occurrence of d_n in that training image. However the likelihood that not a single exact replicate of the data event d_n can be found increases dramatically with the size n of that data event. Indeed, for an attribute S taking K possible states, the total number of possible data events associated with a given n-data template τ_n is K^n (for n=50 and K=2, $K^n \approx 10^{15}$!), while the total number of data events associated with τ_n scanned in the training image is always necessarily smaller than the size N_{TI} of that training image (typically less than a few millions nodes).

When no occurrence of a conditioning data event d_n is present in a training image, the solution proposed by Guardiano and Srivastava (1993), and implemented in the original **snesim** program, consists of dropping the farthest away conditioning data until at least one training replicate of the resulting smaller conditioning data event can be found.

However, *n'* being the number of conditioning data actually used to estimate the conditional probability distribution at the unsampled node **u**, critical information, in particular large-scale information, may be ignored when dropping the (*n-n')* farthest away conditioning data. Such approximation may lead to the inaccurate estimation of some conditional facies probability distributions, and the subsequent simulation of facies values that may not be consistent with the information carried by the dropped conditioning data.

To illustrate the above explanation for the presence of anomalies in MPS models, we plotted in Figure 2b the locations of the nodes that were simulated using less than 10 conditioning data in the MPS model created in the previous section (Figure 1b, repeated in Figure 2a). As expected, a close correspondence can be observed between the locations of those poorly-conditioned nodes and the channel disconnection occurrences.

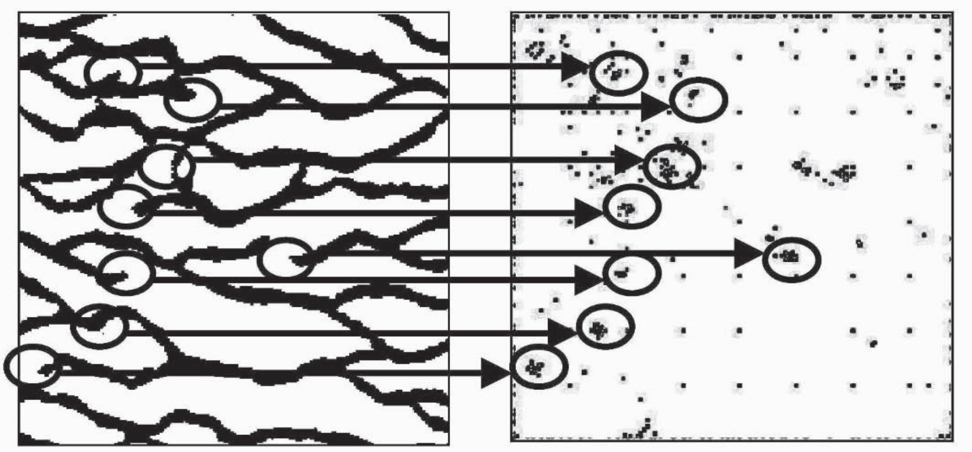


Figure 2. a) MPS model of a 2D horizontal section of a fluvial reservoir generated using **snesim** (left); b) Locations of the nodes simulated using less than 10 conditioning data (right). The arrows show the correspondence between the main clusters of poorly-conditioned nodes and the channel disconnections observed in the MPS model.

In the next section we propose modifying the original **snesim** implementation to decrease the number of dropped conditioning data.

3 Enhanced method to infer conditional probabilities

In the original **snesim** program, the facies probability distribution conditional to a data event $d_n = \{S(\mathbf{u}_1)=s(\mathbf{u}_1)\ldots S(\mathbf{u}_n)=s(\mathbf{u}_n)\}$ (or $d_n = \{s(\mathbf{u}_1)\ldots s(\mathbf{u}_n)\}$ to simplify the notations), is estimated using the following process:

- Retrieve from the search tree the number $c(d_n)$ of d_n-replicates that can be found in the training image.
- If $c(d_n)$ is greater or equal to 1, identify the conditional facies probabilities as the facies proportions of type (1). Otherwise drop the farthest away conditioning datum, reducing the number of conditioning data to (n-1).

Retrieve again from the search tree the number of training replicates of that lesser data event $d_{n-1}=\{s(\mathbf{u_1})\ldots s(\mathbf{u_{n-1}})\}$, and so on… until at least one replicate of the sub-data event $d_n=\{s(\mathbf{u_1})\ldots\ s(\mathbf{u_{n'}})\}$ ($n'\leq n$) can be found in the training image.

- If the number of conditioning data drops to 1, and still no training replicate of d_1 can be found, then the conditional facies probabilities are identified as the target marginal facies proportions of the simulation.

Instead of starting from the full data event d_n, and dropping conditioning data, one can obtain the exact same result using a reverse process, starting from the smallest possible sub-data event d_1, and adding conditioning data until the corresponding conditional probability distribution (cpdf) cannot be inferred anymore from the training image. In more detail, that inverse process consists of the following steps:

- Retrieve from the search tree the number of replicates of the sub-data event $d_1=\{s(\mathbf{u_1})\}$, consisting of a single conditioning datum, that closest to the node **u** to be simulated.
- If no training replicate of d_1 can be found, the local conditional facies probabilities are identified as the target marginal facies proportions of the simulation. Otherwise retrieve again from the search tree the number of replicates of the larger sub-data event $d_2=\{s(\mathbf{u_1}),s(\mathbf{u_2})\}$, consisting of the two conditioning data closest to **u**.
- If no training replicate of d_2 can be found, the probability distribution conditional to d_1 is used to simulate **u**. Otherwise retrieve from the search tree the number of replicates of the larger sub-data event d_3 consisting of the three conditioning data closest to **u**, and so on… until at least one replicate of the sub-data event $d_{n'}$ ($n'\leq n$) can be found in the training image, but no replicate of $d_{n'+1}$.

Because the dropped conditioning data may carry critical information, especially information about large-scale training structures, we propose extending the previous reverse process to retain additional conditioning data beyond $s(\mathbf{u_{n'}})$:

- Given that no replicate of $d_{n'+1}=\{s(\mathbf{u_1})\ldots s(\mathbf{u_{n'+1}})\}$ can be found in the training image, drop $s(\mathbf{u_{n'+1}})$, but add the next conditioning datum $s(\mathbf{u_{n'+2}})$. Retrieve from the search tree the number of replicates of the resulting sub-data event $\{s(\mathbf{u_1})\ldots s(\mathbf{u_{n'}}),s(\mathbf{u_{n'+2}})\}$ (or $d_{n'+2}-\{s(\mathbf{u_{n'+1}})\}$)
- If no training replicate of the previous sub-data event can be found, drop $s(\mathbf{u_{n'+2}})$, otherwise keep that conditioning datum. In both cases, consider the resulting data event, and add the next conditioning datum $s(\mathbf{u_{n'+3}})$, and so on… until the last conditioning datum $s(\mathbf{u_n})$ is reached.

This new conditional probability distribution function (cpdf) estimation method enables the retention of more conditioning data, as confirmed by its application to the previous fluvial reservoir section: only 127 nodes were simulated using less than 10 conditioning data in the new MPS model (Figure 3d) versus 285 in the original one (Figure 2b, repeated in 3c). In particular, the additional conditioning data used are located farther away from the node to be simulated. Therefore large-scale information was better integrated in the new MPS model (Figure 3b) than in the original one (Figure 1b, repeated in Figure 3a), and a significant number of channel disconnections were removed. A post-processing technique is proposed in the next section to remove the remaining anomalies.

Note also that the new cpdf estimation method requires only a minor amount of additional cpu-time: generating a simulated realization using the new cpdf estimation method took 16.3 seconds versus 16.0 seconds for the original simulated realization on a 660MHz SGI Octane II.

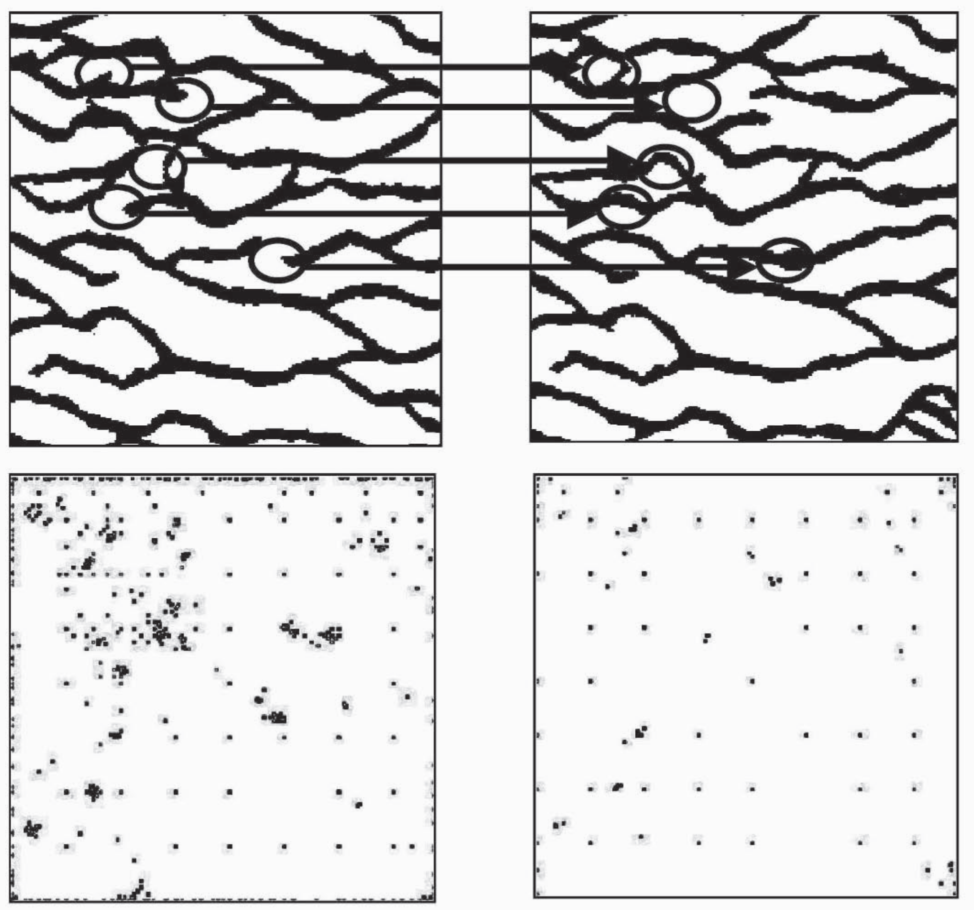


Figure 3. MPS models generated using a) the original cpdf estimation method (top left); b) the new cpdf estimation method (top right). Locations of the nodes simulated using less than 10 conditioning data c) in the original MPS model (bottom left); d) in the new MPS model (bottom right). The arrows show the channel disconnections that were removed from the original model.

4 A new post-processing algorithm

A post-processing algorithm was proposed by Remy (2001) to remove anomalies from MPS models using a two-step process:

- First, given a data template τ_N, identify in the simulated realization all data events associated with τ_N that do not occur in the training image
- Then re-simulate the grid nodes of those data events.

However, because of the limited size of the template τ_N, only small-scale anomalies could be corrected. Furthermore, a better identification of the nodes to be re-simulated is proposed in this section.

When estimating local cpdf's, conditioning data are dropped until at least one replicate of the resulting conditioning data event can be found in the training image. Considering a larger training image could provide additional possible patterns, thus decreasing the number of dropped conditioning data. But in many cases, a datum actually must be dropped because the information it carries is not consistent with the information carried by the other nearby conditioning data. Dropped conditioning data may be then a good indicator of the local presence of anomalies. This is confirmed by the good spatial correlation observed between the channel disconnections of the previous fluvial reservoir MPS model (Figure 3b, repeated in Figure 4a) and the clusters of nodes where conditioning data were dropped (Figure 4b).

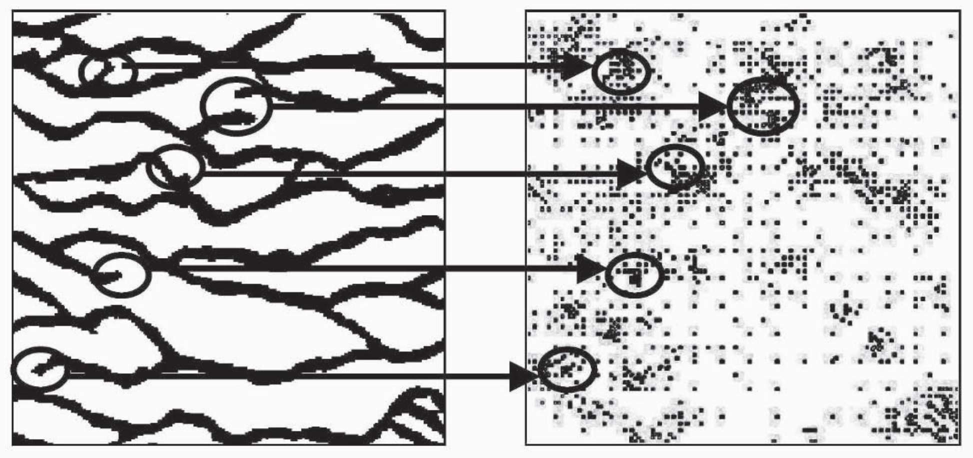


Figure 4. a) MPS model obtained using the new cpdf estimation method proposed in the previous section (left); b) locations of the conditioning data dropped (right). The arrows show the correspondence between channel disconnections and clusters of dropped conditioning data.

We propose marking the locations of the conditioning data dropped as the simulation progresses from one grid node to the other and revisiting these locations later. The new **snesim** implementation using that post-processing method proceeds in the following steps:

1. Define a random path visiting once all unsampled nodes.

2. At each unsampled node **u**, retrieve the local conditional probability distribution using the new cpdf estimation method previously described, and mark the nodes corresponding to the conditioning data dropped. Draw a simulated value for node **u**.
3. Move to the next node along the random path and repeat step 2.
4. Once all grid nodes have been visited and simulated, remove values from the nodes that have been marked, provided that they do not correspond to original sample data.
5. Repeat steps 1 to 4 several times until the generated image is deemed satisfactory according to some convergence criterion, for example until the number of nodes to be re-simulated stops decreasing.

To correct anomalies at all scales, this post-processing technique needs to be applied to every nested (fine or coarse) grid used in the multiple-grid simulation approach implemented in **snesim**. Figure 5 shows two post-processed MPS models of the previous fluvial reservoir. The number of channel discontinuities in both models is much lower than in the original model of Figure 4a.

Six post-processing iterations were performed on average per nested grid, and 43 nodes were re-simulated on average per iteration. The additional cpu-time required for the post-processing is relatively small: generating one realization using post-processing took 19.2 seconds on average, versus 16.3 seconds without post-processing.

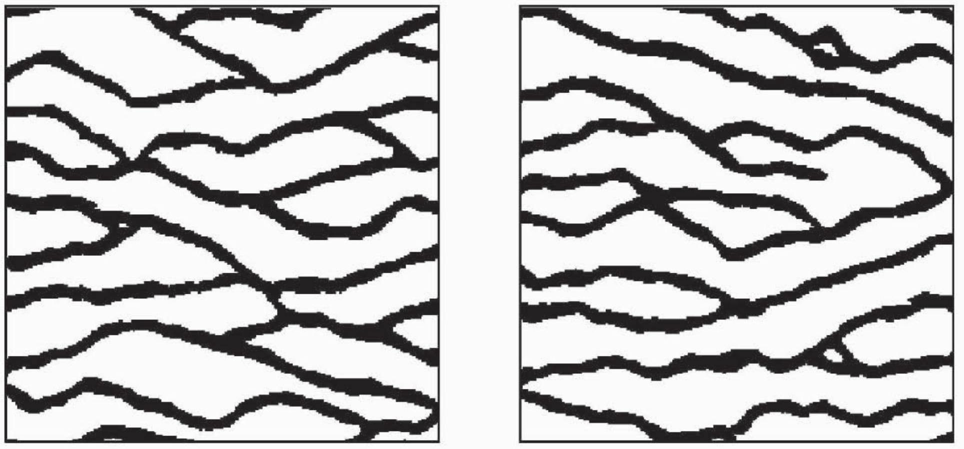


Figure 5. Two MPS models generated using the post-processing.

In 3D applications, the number of nodes to be re-simulated may be much higher than in 2D. Thus, instead of simply marking the nodes where conditioning data were dropped, we propose measuring the local consistency of the MPS model with regard to the training image by counting the number of times each simulated value, although being initially identified as a conditioning datum to simulate a nearby node, had to be ignored eventually to be able to infer from the training image the conditional probability distribution at that node. The post-processing consists then of re-simulating only those nodes where the previous consistency measure is greater than a given threshold.

5 Conclusion

A new version of the multiple-point statistics simulation program **snesim** with integrated post-processing is presented in this paper. In this new program, the method used to infer local conditional facies probability distributions is modified to increase the number of conditioning data actually used in that inference process. This new estimation method removes from multiple-point geostatistical models a great number of anomalies, i.e. simulated patterns that were not present in the training image. Then a post-processing technique is proposed to reduce the number of remaining anomalies. The application of that post-processing to a 2D horizontal section of a fluvial reservoir shows that the cpu-time needed to run the new modified **snesim** is comparable with that of the original **snesim**, while the number of anomalies decreases dramatically.

References

Deutsch, C.V. and Journel, A.G., *GSLIB: Geostatistical Software Library and User's Guide*, 2nd edition, Oxford University Press, 1998.

Guardiano, F. and Srivastava, R.M., Multivariate Geostatistics: Beyond Bivariate Moments, in Soares, A., editor, *Geostatistics-Troia*, vol. 1, p. 133-144. Kluwer Academic Publications, 1993.

Remy, N.. Multiple-point statistics for image post-processing, in *Report 14, Stanford Center for Reservoir Forecasting*, 2001.

Strebelle, S., *Sequential Simulation Drawing Structures from Training Images*, Ph.D. Thesis, Department of Geological and Environmental Sciences, Stanford University, 2000.

Tran, T., Improving Variogram Reproduction on Dense Simulation Grids, *Computers and Geosciences***,** vol. 20, no. 7, 1994, p. 1161-1168.

Viseur, S., Stochastic Boolean Simulation of Fluvial Deposits: a New Approach Combining Accuracy and Efficiency, paper SPE 56688 presented at the 1999 SPE Annual Technical Conference and Exhibition, Houston, Oct. 3-6, 1999.

IMPROVING THE EFFICIENCY OF THE SEQUENTIAL SIMULATION ALGORITHM USING LATIN HYPERCUBE SAMPLING

GUSTAVO G. PILGER, JOÃO FELIPE C. L. COSTA and JAIR C. KOPPE
Mining Engineering Department,
Federal University of Rio Grande do Sul,
Av. Osvaldo Aranha 99/504 90035-190,
Porto Alegre, Brazil

Abstract. Sequential simulation is probably the most used algorithm in geostatistical simulation, specially the parametric version, i.e. the sequential Gaussian algorithm. This algorithm requires the data to follow a Gaussian distribution, and assumes also that all multivariate distributions are also normal. This assumption is very convenient allowing the local uncertainty model (or conditional cumulative distribution function - ccdf) to be inferred through few parameters given by simple kriging (mean and variance). The set of simulated values are drawn through Monte-Carlo methods, randomly sampling the ccdf L times. In theory, this method maps the space of uncertainty as the number of realizations increase. In practice the number of realizations necessary varies according to the characteristics of the conditioning data. It is important that the L simulations describe the space of the uncertainty appropriately accordingly to the objective addressed. However, in some situations the number of simulations needs to be large, making the procedure computationally intense and time-consuming. This paper presents a more efficient strategy to generate the local ccdf based on Latin Hypercube Sampling (LHS) technique. The idea is to replace the Monte-Carlo simulation by the LHS in order to improve the efficiency of the sequential Gaussian simulation algorithm. The use of the modified algorithm showed that the space of uncertainty related to the random variable modeled was obtained faster than the traditional Monte-Carlo simulation for a given degree of precision. This approach also ensures that the ccdf is better represented in its entirety.

1 Introduction

Geostatistical simulation applications in the mining industry are quickly growing. These methods have been employed for risk analysis during pre-feasibility through feasibility studies, as in mine planning (short and medium term) and production stages. Most applications are aimed at sensitivity analysis on grade-tonnage curves and their effect on the projects net present value. Stochastic conditional simulation is able to quantify the variability on geological attributes such as grades or any other variable relevant to a given mining project. Variability can be assessed by constructing multiple equally probable numerical scenarios. Combining these scenarios can provide an assessment on the so called space of uncertainty.

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 989–998.

Geostatistical simulation methods normally are set to create L (L = 1,...,L) equally probable images of any attribute from a mineral deposit, reproducing their 1st. and 2nd. order statistical moments. To each point or block of a simulated domain there are L equally probable values of the attribute. The variability of these simulated values can be assessed by uncertainty indices. These indices are calculated after the simulation process, using the L equiprobable values. Among the uncertainty indices one can use:

i. conditional variance;
ii. conditional coefficient of variation;
iii. interquartiles (interquantiles) ranges;
iv. entropy.

Thus, it is possible to verify the amplitude of variation among several equally probable scenarios and calculate the error associated with the estimates. It is assumed that the basic statistics and the spatial dependence derived from the samples are representative of the parent population.

The most used stochastic conditional simulation algorithms are the sequential Gaussian (Isaaks, 1990), sequential indicator (Alabert, 1987) and the turning bands method (Matheron, 1973). These algorithms are available in the most geostatistical softwares such as GSLIB (Geostatistical Software's Library) (Deutsch and Journel, 1998) or Isatis®. Amongst the cited methods, the sequential ones, parametric or nonparametric, are preferentially used.

The main difference between these two groups is the procedure used for constructing the uncertainty models (conditional cumulative distribution function - ccdf): parametric vs. nonparametric. Sequential Gaussian simulation (sGs) is based on the multiGaussian formalism (parametric), while the sequential indicator simulation (sis) uses the homonym formalism (nonparametric).

The multiGaussian approach assumes that all multivariate distribution of the data follows a Gaussian distribution. Thus, the application of sGs algorithm demands the experimental distribution of the random variable (RV) Z(u) follows a Gaussian distribution. That is, the RV Z(u) must be transformed into a RV Y(u) standard normal. The multiGaussian hypothesis is very convenient, as it allows the uncertainty models (ccdf) to be obtained from a normal distribution with mean and variance derived from kriging. Thus, the mean and the variance of the ccdf in a given unsampled location, u, are equal to the respectively estimate $y_{SK}^{*}(u)$ and variance $\sigma_{SK}^{2}(u)$ of simple kriging (SK). Then, the ccdf can be modeled as:

$$\left[G(u;y|(n))\right]_{SK}^{*} = G\left(\frac{y - y_{SK}^{*}(u)}{\sigma_{SK}(u)}\right) \tag{1}$$

where y is a Gaussian value of the domain [-∞; +∞]. The estimated values $y_{SK}^{*}(u)$ and $\sigma_{SK}^{2}(u)$ are calculated from n information $y(u_\alpha)$ ($\alpha = 1,...,n$) in the neighborhood of u (Journel and Huijbregts, 1978 p. 566).

After constructing the ccdf, a simulated datum $y^{(l)}(u_j)$ is draw from it via Monte-Carlo simulation. Generally, the following stages are common to all stochastic sequential simulation algorithms (parametric or nonparametric):

i. definition of a random path, in which each unsampled location u_j ($j = 1...,N$) (point, cell or block of the grid) is visited only once;
ii. construction of the uncertainty model (ccdf) at the location u_j – conditional to the n experimental information in the neighborhood of u_j;

iii. simulation of a value $y^{(l)}(u_j)$ from the RV $Y(u_j)$, by drawing randomly from the ccdf (Monte-Carlo simulation);

iv. inclusion of $y^{(l)}(u_j)$ into the data set, representing an addition to the conditional information to be used in the following N grid nodes to be visited $\{y^{(l)}(u_j), j = 1...,N\}$;

v. repetition of the stages (ii) to (iv) until a simulated value is associated to each of the N locations;

vi. repetition of the steps (i) the (v) to generate L equiprobables realizations of the spatial distribution of the RV Y(u).

Hence, the set $\{y^{(l)}(u_j), j = 1,...,N\}$ represents a realization of the random function (RF) Y(u) in the physical domain defined by the information $y(u_\alpha)$ ($u_\alpha = 1,...,n$), in the normal space. Whereas the set $\{y^{(l)}(u_j), l = 1,...,L\}$ represents L simulations of the RV Y at location u_j (j = 1,...,N). After, the simulated data set $\{y^{(l)}(u_j)$ (j = 1,...,N and l = 1,...,L)$\}$ is transformed to the original space of the RV Z(u). Therefore, the value of the RV Z at each location u_j (j = 1,...,N) is simulated within the domain of variation of the RV Z(u), through a random procedure, from the ccdf. At each location, the simulation process generates a distribution, composed by L values. That distribution can be considered a numerical approach of the ccdf; i.e.:

$$F(u;z\,|\,(n)) \approx \frac{1}{L}\sum_{l=1}^{L} i^{(l)}(u;z) \qquad (2)$$

where F(u; z| (n)) represents the probabilities assumed by the ccdf at each location u_j (j = 1,...,N) and $i^{(l)}(u;z)$ is a indicator variable as follows:

$$i^{(l)}(u;z) = \begin{cases} 1 \quad if \quad z^{(l)}(u) \le z & with \quad l = 1,...,L \\ 0 \quad if\ not \end{cases}$$

Indeed, the methods of stochastic sequential simulation make use of a very curious algorithm. The objective is to obtain a model of uncertainty, however this model is known beforehand, as after kriging all the parameters of the multiGaussian distribution are determined.

Generally, the stochastic methods use the technique of Monte-Carlo simulation (Isaaks, 1990) to construct the numerical models. The model of uncertainty (ccdf) of the RV Y at u_j (j = 1,...,N) locations is randomly sampled L times, generating the set of simulated values $\{y^{(l)}(u_j)$ (j = 1,...,N and l = 1,...,L)$\}$ and, consequently the set $\{z^{(l)}(u_j)$ (j = 1,...,N and l = 1,...,L)$\}$. In theory, the stochastic methods reproduce the space of uncertainty of the RV Z(u) as the number L of realizations increases. However, in practice, this number is a function of the experimental distribution, requiring to be as large as the variability of data. The number L of random drawings should be sufficiently large to guarantee the space of uncertainty of the RV Z(u) is characterized.

However, frequently L needs to be so large that the technique becomes too computationally intense. The dimension of the problem (2D or 3D), the size of the area to be evaluated (number of points or blocks on the grid), the number of conditioning data, and their statistical characteristics (auto-correlation or variability) can be barriers for the adequate use of the stochastic simulation. Various case studies exist to corroborate the problem (Godoy, 2002; Santos, 2002). Koppe *et al.* (2004) simulated in 3D sonic wave velocity obtained by geophysical logging. Given the amount of conditioning data and the size of the grid, many hours of CPU were necessary (Pentium 4, 2.8 MHz, 1024 Mb) to generate a single realization. Many industry applications require prompt answers and do not allow the processing of an adequate number of realizations before a decision is made. Although the use of geostatistical simulation is increasing among the mining industry, the correct application of this technique in certain situation can be difficult due to the reasons listed above (related to the characterization of the space of uncertainty). Thus, in these situations the variability of the studied mineral attribute will not be correctly evaluated.

This paper shows a more efficient strategy able to cover with fewer realizations the space of uncertainty. The Sequential Gaussian simulation algorithm was modified, aiming at increasing its efficiency. The replacement of Monte-Carlo simulation by Latin Hypercube Sampling technique (LHS) (McKay *et al.*, 1979) is proposed. The proposal attempts to build the so-called space of uncertainty of the RV Z(u), for a given precision, with fewer runs.

2 Methodology

Stochastic simulation requires a minimum number of realizations, L, to ensure the space of uncertainty of the RV Z(u) is characterized. The substitution of the random drawing embedded in Monte-Carlo simulation for LHS can increase the efficiency of the process. For a given precision, LHS guarantees that all ccdf is sampled (in it is integrity), generating the set of L simulated values $\{z^{(l)}(u_j)\ j = 1,...,N \text{ and } l = 1,...,L)\}$, with fewer realizations than Monte-Carlo.

The algorithm of sequential Gaussian simulation is implemented to get the simulated value $y^{(l)}(u_j)$ (in the normal space) by means of the following equation:

$$y^{(l)}(u_j) = xp^{(l)}(u_j)\ .\ \sigma_{SK}(u_j) + y^*_{SK}(u_j) \qquad u_j = 1,...,N \text{ and } l = 1,...,L \tag{3}$$

where $\sigma_{KS}(u_j)$ and $y^*_{KS}(u_j)$ represent the simple kriging standard deviation and the simple kriging estimate, respectively at location u_j . For each new l, or new run, these values change due to the random path used to visit the nodes and the addition of previously simulated nodes to the original dataset. Whereas $xp^{(l)}(u_j)$ represents a Gaussian value, obtained from the standard Gaussian probability distribution function G(y), that is:

$$xp^{(l)}(u_j) = [G^{-1}(u_j, p)] \qquad u_j = 1,...,N \text{ and } l = 1,...,L \tag{4}$$

where p is a probability defined in the domain [0; 1], obtained randomly by means of the Monte-Carlo method.

LHS consists of randomly drawing, without substitution (stratified random sampling without substitution), of M values, from M distinct equally probable classes of a ccdf. Thus, in case of using LHS instead of Monte-Carlo, M values are drawn randomly and without substitution from the standard Gaussian probability distribution function G(y) representing the standard Gaussian RV Y (Y~N(0,1)). However, LHS requires the M values be drawn from M distinct equally probable classes. Firstly, the G(y) distribution is split into M disjunctive equally probable classes. From there, values are randomly drawn from each class m (m = 1,…,M), at each location u_j (j = 1,...,N) for each L (l = 1,...,L) realization. The drawing procedure follows:

$$xp_{lhs}^{(m,l)}(u_j) = [G^{-1}(u_j, (plhs^{(m,l)})] \qquad u_j = 1,...,N; \; m = 1,...,M \text{ and } l = 1,...,L \qquad (5)$$

where $plhs^{(m,l)}$ (a probability within a class m, among the range of M classes) is defined as:

$$plhs^{(m,l)}(u_j) = ((m-1) + R_m)/M \qquad u_j = 1,...,N; \; m = 1,...,M \text{ and } l = 1,...,L \qquad (6)$$

where R_m is a random number defined in the domain [0; 1].

Thus, at each location u_j (j = 1,...,N), a Gaussian value $xp_{lhs}^{(m,l)}$ is randomly taken from each class M. Note that to ensure that one Gaussian value $xp_{lhs}^{(m,l)}$ is drawn from each class M, the number of realizations L (random drawings) must to be equal the number of classes M used to discretize the G(y) distribution, i.e. L = M.

Following a random path, each location u_j (j = 1,..., N) is visited and a SK system is solved to obtain $\sigma_{SK}(u_j)$ and $y^*_{SK}(u_j)$. Next, one class M is randomly drawn (without substitution) and a probability ($plhs^{(m,l)}(u_j)$), defined in that class, is determined by solving Equation 6. Next, a Gaussian value $xp_{lhs}^{(m,l)}(u_j)$ is obtained (Equation 5). Finally, the simulated values $y^{(m,l)}(u)$ (m = 1,..., M and l = 1,..., L) are obtained as follow:

$$Y^{(m,l)}(u_j) = xp_{lhs}^{(m,l)}(u_j) \, . \, \sigma_{SK}(u_j) + y^*_{SK}(u_j) \qquad u_j = 1,..., N; \; m = 1,..., M \text{ and } l = 1,...,L \qquad (7)$$

It is important to stress that m is randomly taken, and it can assume any integer within the interval [1; M]. Conversely, the values for l are sequential within the [1; L] interval. In the proposed algorithm, the user has to inform the usual parameters of sGs, and additionally has to define the number M of classes (i.e. define the way G(y) is discretized). Therefore, the new stochastic sequential simulation algorithm would be:

i. define M disjunctive equally probable classes;
ii. define a random path, in which each unsampled location u_j (j = 1,..., N) (cell or block of the grid) is visited once;
iii. solve a kriging system (SK) to estimate $\sigma_{SK}(u_j)$ and $y^*_{SK}(u_j)$ at each visited location u_j, giving $n(u_j)$ experimental (and previously simulated data) within the neighborhood of u_j;
iv. random sample without substitution the class defined in (i);
v. random sample the probability $plhs^{(m,l)}$ (Equation 6);
vi. draw a Gaussian value $xp_{lhs}^{(m,l)}$ (Equation 5);
vii. simulate a value $y^{(m)}(u_j)$ (Equation (7);
viii. add $y^{(m)}(u_j)$ to the dataset;
ix. repeat steps (iii) to (viii) until a simulation is associated to each one of the N locations;
x. repeat steps (ii) at (ix) to generate L (L = M) equally probable realizations of the RF Y(u);

xi. back transform the normalized values $y^{(m)}(u_j)$ to obtain the realizations of $Z(u)$.

This modification adds a new random component into the original algorithm. In addition to drawing the probabilities ($plhs^{(m,l)}$) there is also the drawing (without substitution) of equally probable disjunctive classes M.

Figure 1 shows the experimental Gaussian probability distributions generated by Monte-Carlo and LHS. These distributions were obtained by plotting 20 outcomes for a location u_j ($j = 1$), calculated by equations (4) and (5). The Gaussian values ($xp^{(l)}$ and $xp_{lhs}^{(m,l)}$) are plotted against their probabilities. LHS better reproduces the G(y) distribution, especially on its extremes and the stratification is evident in the LHS plot (Figures 1 and 2). Consequently, the ccdf of the RV Z(u) is expected to be better constructed, i.e. sampled uniformly at all classes (although there is no guarantee that the resulting ccdf of Z(u) is stratified). In this case L = 20 was used; therefore to use LHS at location u_j ($j = 1$), 20 random draws were taken from 20 classes with 5% of amplitude each.

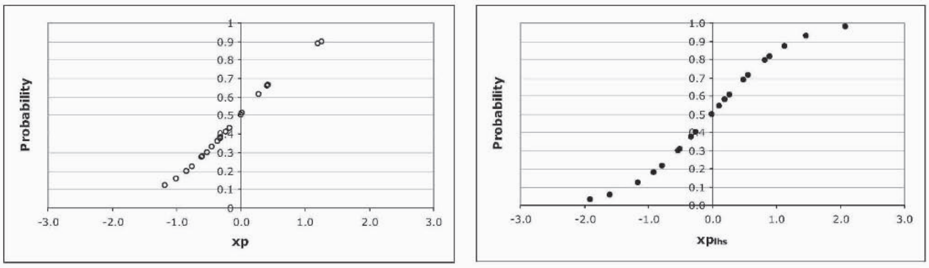


Figure 1: Twenty outcomes (realizations) of a Gaussian distribution sampled by Monte-Carlo (left) and by LHS (right).

Figure 2 presents the probabilities related to same 20 outcomes (L = 20) and corroborates the results verified in Figure 1. The horizontal axis plots the realization number while the vertical axis the respective drawn probability. A straight line in this plot would indicate that the G(y) distribution was properly sampled. LHS plots closer to a straight line than the results obtained by Monte-Carlo.

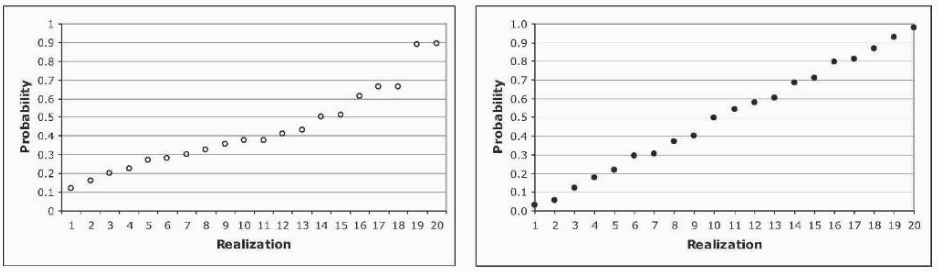


Figure 2: Probabilities drawn for 20 realizations via Monte-Carlo (left) and via LHS (right).

A similar comparison was conducted for several distinct values of L. LHS consistently presented better reproduction of the parent Gaussian distribution than Monte-Carlo, independently of the location u_j ($j = 1,..., N$) chosen.

3 Discussion on the results

In order to evaluate practical benefits on the modified sGs algorithm, several tests were performed and compared to the results obtained using the original sGs. For these tests, the Walker Lake dataset (Isaaks and Srivastavas, 1989) was used and its parent population is available.

Two hundred simulation were generated (L = 200) using the original sGs algorithm. The variability on the mean of the realizations was calculated using the coefficient of variation (CV). The average and CV of the mean of the realizations were calculated for an increasing set of realizations. That is, the 10 first realizations had been analyzed, next the 20 first ones and so on. For instance, the initial set C of the 10 first realizations (C = 10) was analyzed and the average of the mean of all 10 realizations ($E[m]_{10}$) and the respective $CV(m)_{10}$ of this set had been calculated.

Figure 3(a) plots the average of the mean of several sets of realizations. The average reaches a stable value after the 160th realization. Figure 3(b) presents the CV of the means for the same realization sets. Similarly, the CV stabilizes after the 150th realization.

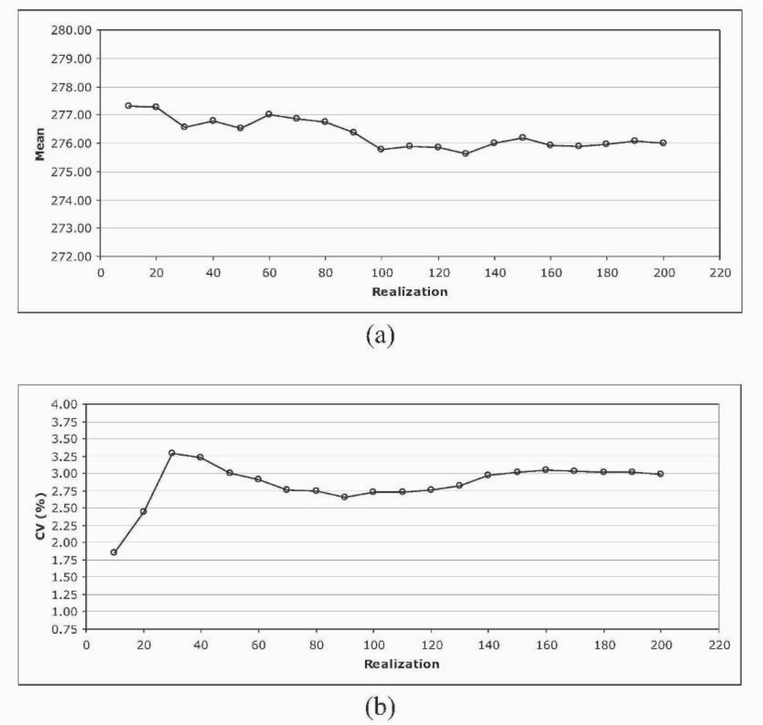


Figure 3**:** Average (a) and CV (b) of the mean of the realizations using original sGs.

The difference between the average of the mean values in relation to the sets of 100 and 160 realizations is small, however the CV is significantly different. Generally, users of simulation tend to set L equal 100 to guarantee the space of uncertainty of the RV Z(u) is properly covered, ignoring the statistical characteristics of the data. However, this example highlights the fact that in some situation 100 realizations is not enough (or in other cases is too much) to assess the space of uncertainty of the RV Z(u).

The procedure was repeated using the modified sGs algorithm. Several class amplitudes were tested. The results were satisfactory, even for a small number of realizations (few classes, as L = M). The modified sGs algorithm convey a better characterization of the space of uncertainty of the RV Z(u). Figures 4 and 5 compare the results obtained for the two algorithms. For these figures, the number of simulations added on each step was one (C = 1). The steady values for average and CV are also plotted, i.e. the values which the two statistics become stable. Figure 4 shows that after 14 realizations the average of the realization's mean reaches a stable value. LHS in this situation required the standard Gaussian probability distribution G(y) to be divided into 14 classes (M = 14) with an amplitude of 7,14% each class. The CVs of these means are presented in Figure 5. In relation to the variability of the mean for sets of realizations, the performance of the modified sGs is also superior. Figures 4 and 5 also show that using M > 14 the statistics still fluctuate around the steady values, however the results are better than the original sGS algorithm. The results obtained for M > 20 are similar to those obtained by the original sGS algorithm. (or: The modified algorithm becomes similar to the original sGS algorithm for M > 20).

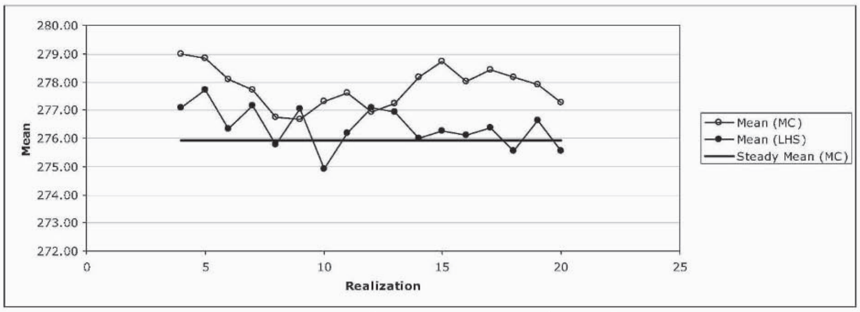


Figure 4: Average for the mean of the realizations.

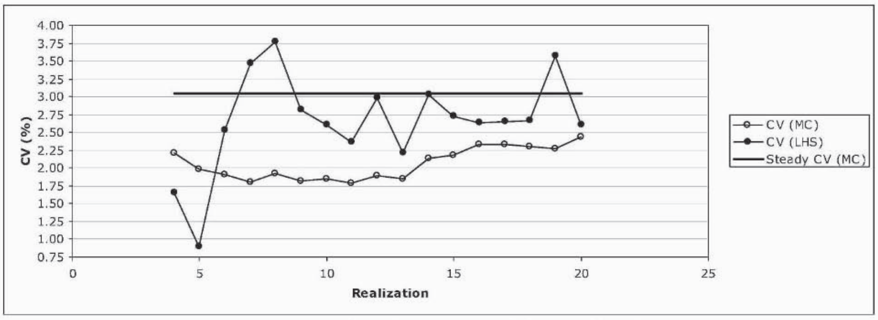


Figure 5: CV for the mean of the realizations.

From a practical perspective, the impact of underestimating the size of the space of variability is analyzed via recoverable reserves curves of a mineral deposit. Figure 6 illustrates the cutoff vs. tonnage curve for 14 equally probable scenarios generated by the two methods.

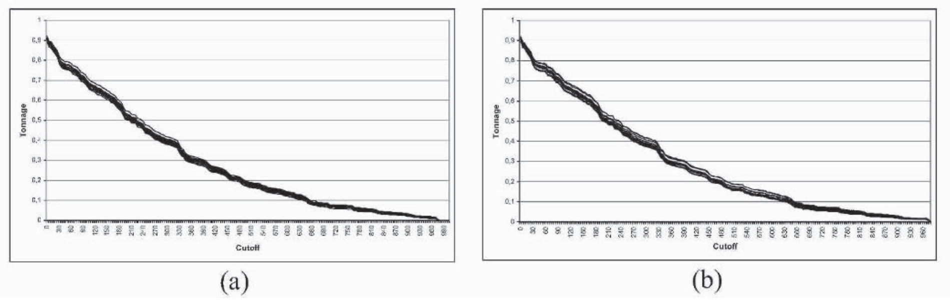


Figure 6: Recoverable reserves curves (cutoff x tonnage) obtained for 14 realizations obtained by original sGs (a) and by the modified sGS (b).

Although the two sets of curves seem very similar, the amplitude of variation (spread) of the curves produced by the 14 realizations generated by the modified sGs is larger (Figure 7). The modified algorithm is more conservative in the sense the space of uncertainty is wider, as shows the Figure 8.

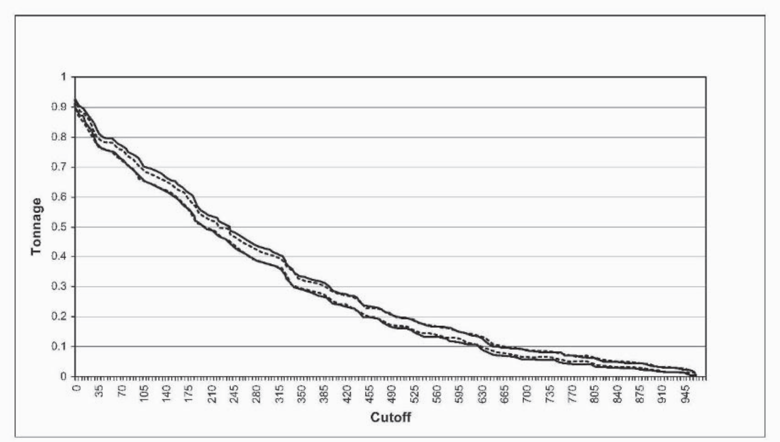


Figure 7: Amplitude of the recoverable reserves curves (cutoff x tonnage), original sGs (dashed line) and modified sGs (solid line).

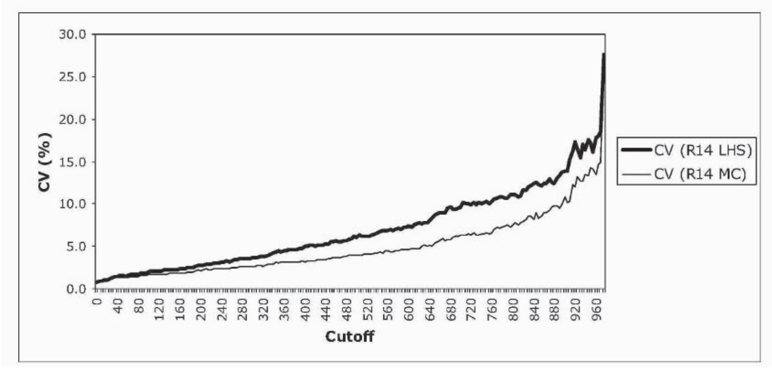


Figure 8: CV for the 14 values of tons (obtained from the realizations) calculated for original sGS (thin line) and modified sGs (thick line).

Figure 8 presents the CV for the tons calculated at various cutoffs using the 14 possible outcomes obtained by simulation. This figure reflects the previous conclusions, i.e. the variability on possible values for tonnage is higher (more uncertainty) using the modified algorithm. It is expected that using the original algorithm it will be necessary more than 100 realizations to achieve the same variability.

4 Conclusion

The number of realizations necessary to adequate characterize the variability of the mineral attributes can significantly be reduced. Thus, the adequate use of the geostatistics simulation methods in the industry can be facilitated by generating smaller output files. Hence, the L number of simulation can be reduced (for the same space of uncertainty obtained by the original sGs) helping to handle smaller output file.

The modified algorithm produced a more conservative answer, i.e. a larger space of uncertainty. Therefore, in industry applications where the number of realizations is a limiting factor, the modified sGs algorithm can be more appropriate, as with fewer realizations the variability (space of uncertainty) achieved is similar to the one obtained via original sGs. The proposed method does not ensure that the resulting ccdf are stratified, but the improvement in mapping the consequential transfer function's space of uncertainty is visible as demonstrated in the case study.

In practice, the challenge is to find the number M of classes (that discretize the G(y) distribution) adjusted to each situation. The choice of M is still a subject to be studied. The reproducibility of the results also needs to be verified on a few other datasets.

References

Alabert, F., *Stochastic Imaging of Spatial Distributions Using Hard And Soft Information,* M.Sc. Thesis, Stanford University, Stanford, 197p, 1987.

Deutsch, C.V. and Journel, A.G., *GSLIB: Geostatistical Software Library and User's Guide*, 2nd Edition, Oxford University Press, New York, 345p, 1998.

Godoy, M. C., *The Effective Management of Geological Risk in Long-Term Production Scheduling of Open Pit Mines*, PhD. Thesis, The University of Queensland, Australia, 256p, 2002.

Isaaks, E., H., *The Application of Monte Carlo Methods to the Analysis of Spatially Correlated Data*, Ph.D. Thesis, Stanford University, USA, 213p, 1990.

Journel, A. G. and Huijbregts, C. J., *Mining Geostatistics*, Academic Press, London, 600p, 1978.

Koppe, V. C.; Gambin, F.; Costa, J.F.C.L.; Koppe, J.C.; Fallon, G. and Davies, N., Análise de Incerteza Associada à Determinação da Velocidade de Onda Sônica em Depósitos de Carvão Obtida por Perfilagem Geofísica – Newland Coal Mine – Austrália, *in III Congresso Brasileiro de Mina A Céu Aberto & III Congresso Brasileiro de Mina Subterrânea,* Belo Horizonte, MG, 2004.

Matheron, G., The Intrinsic Random Functions and Their Applications, *Advances in Applied Probability*, vol. 5, 1973, p. 439-468.

McKay, M. D.; Beckman, R. J. and Conover, W. J., A Comparison of Three Methods For Selecting Values of Input Variables in The Analysis of Output From a Computer Code. *Technometrics*, vol. 21, 1979, p. 239-245.

Santos, Sérgio Fernando dos, Redução da Incerteza Associada a Modelos Estocásticos de Fácies Através do Condicionamento a Dados de Produção., MSc Dissertation, Universidade Federal do Espírito Santo, 120p, 2002.

EXACT CONDITIONING TO LINEAR CONSTRAINTS IN KRIGING AND SIMULATION

J. JAIME GÓMEZ-HERNÁNDEZ
Universidad Politécnica de Valencia, Spain

ROLAND FROIDEVAUX
FSS Consultants SA, Geneva, Switzerland

PIERRE BIVER
Total, Pau, France

Abstract. A recurrent problem in reservoir characterization is the need to generate realizations of an attribute conditioned not only to core data measurements, but constrained also to attribute averages, defined on a larger support. A related problem is that of downscaling, in which there is a need to generate realizations at a scale smaller than the available data, yet preserving some type of average relation between the data and the downscaled realization. Typical larger support data are those coming from well tests, production data, or geophysical surveys. A solution is proposed to approach these two problems for the case in which the large support data can be expressed as linear functions of the original, smaller support, attribute values, or of some local transform of them. The proposed approach considers two random functions, one for the point (small support) data, and one for the block (large support) data. The algorithm is based on the full specification of the point to block and block to block covariances from the point to point covariance. Once all direct- and cross-covariances are specified, co-kriging or co-simulation can be used to either produce estimation or simulation maps. Similar approaches have been attempted, but this approach distinguishes itself because it is exact, in the sense that the constraints are exactly honored in the final maps. Although the theoretical basis for this constrained estimation or simulation is reasonably straightforward, its implementation is not. In particular, the building of the co-kriging systems and the concept of search neighborhood presents some non-negligible challenges, which have been efficiently solved, even for the non-trivial case of overlapping supports of the constraining data.

1 Introduction

Reservoir or aquifer characterization always faces the problem of how to handle data taken at different supports. When the relationship between the attribute

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 999–1005.

value (or any non-linear transform of it, i.e., the logarithm of the attribute) at different scales is linear, or it can be established in terms of a p-norm (or power average), it is possible to use data at any scale to condition either estimation by kriging or stochastic simulation. Deutsch (1993) approached the problem using simulated annealing, Srinivasan and Journel (1998) used direct sequential simulation to approximately constrain the p-norm average of an entire permeability realization to the well-test derived permeability, and Journel (1999) explained how to approximately constrain kriging and simulation to p-norm averages.

But conditioning can be made exact with the proper implementation based on the fact that the covariance is a measure of the linear relationship between two attributes, thus, covariance reproduction entails linear relationship reproduction. This property was proven by Gómez-Hernández and Cassiraga (2000) who described how to generate realizations conditioned to linear constraints by sequential simulation.

Constraining to linear averages can be important. For instance, geophysical surveys do not provide information on small supports, yet, geophysical measurements can be related to linear averages of an attribute such as porosity in a neighborhood of the geophysics log; or, pumping tests do not provide a measurement of the permeability at the well, but the antilog of a weighted average of the nearby logarithm of the permeabilities;

The methods described by Gómez-Hernández and Cassiraga (2000) in their paper were not suitable neither when the realizations had many cells, nor when the constraints extended over a very large area, for reasons similar to those in Srinivasan and Journel (1998) and Journel (1999) who propose only approximate solutions to the problem of conditioning to average values. In this paper, we describe and demonstrate an algorithm that solves these problems and that can be used for either estimation or simulation.

2 Theory

Consider the random function Z to be simulated over a grid of n_Z points and the random function Y linearly related to Z through

$$A \cdot Z = Y, \tag{1}$$

in which A is an $n_Y \times n_Z$ matrix, with all its rows linearly independent. Random function Z represents the attribute defined at a given support, and Y represents the linear constraints that have to be met by Z.

Since Y is fully determined by Z, its covariance and cross-covariance are fully determined from the covariance of Z, C_Z and (1):

$$\begin{aligned} C_Z &= E\{Z \cdot Z^T\}, \\ C_{ZY} &= E\{Z \cdot Y^T\} = C_Z \cdot A^T, && (2) \\ C_{YZ} &= E\{Y \cdot Z^T\} = A \cdot C_Z, && (3) \\ C_Y &= E\{Y \cdot Y^T\} = A \cdot C_Z \cdot A^T. && (4) \end{aligned}$$

As already mentioned, it has been shown elsewhere that kriging estimation or stochastic simulation of the random function Z can be made exactly conditioned to Y samples if kriging or simulation is performed using the C_Z covariance and the C_Y, C_{YZ}, C_{ZY} covariances given by equations (2-4). Conditioning to y data is equivalent to imposing the corresponding linear constraint onto the realization of Z.

3 Implementation

3.1 KRIGING

When the algorithm is implemented as an extension of standard estimation or simulation packages, the constraints are not exactly reproduced. The mismatch is due to the approximations common to all standard implementations of kriging or simulation.

The single most critical point for the exact reproduction of the constraints is the selection of the set of data retained for kriging estimation at a given point. We cannot renounce to the possibility of using a kriging neighborhood in order to limit the number of points used in the solution of the kriging system; however, when dealing with linear constraints several questions arise:

- When should a constraint be retained as an additional datum?
- If a constraint is retained which expands beyond the search neighborhood, should the search neighborhood be modified?
- How should overlapping constraints be treated?

The proper answers to these questions enable exact reproduction of the linear constraints:

- A constraint datum y_i involves several random function locations $\{Z_j, j \in (n_i)\}$, with (n_i) representing the locations constrained by y_i. If any location in (n_i) is in the search neighborhood of the point being estimated, the constraint y_i must be included in the kriging estimation.
- The search neighborhood must be extended to the union of the initial search neighborhood plus the locations of all the constraints retained according to the previous item.
- If there are overlapping constraints, the previous two steps have to be repeated, until no more extensions of the search neighborhood happen, since new constraint locations may be included in the extended neighborhood.
- All data in the search neighborhood must be retained, not just the closest ones.

The previous considerations imply that the search neighborhood actually used in the estimation will, in general, be larger than the search neighborhood specified by the user. At the limit, if there is one of the constraints that involves all nodes, the search neighborhood will always be global, both for the z and y data. But failing to modify the search neighborhood according to any of the three items above will result in only an approximate reproduction of the constraint.

3.2 SIMULATION

All above considerations must be applied to simulation, too. Gómez-Hernández and Cassiraga (2000) were aware of them and included them all in their proposal of sequential simulation with linear constraints. However, considering that sequential simulation incorporates all simulated nodes as conditioning data, the need to krige using all data values related to the conditioning constraints quickly yields the algorithm unfeasible for constraints involving more than a few tens of cells, or overlapping constraints.

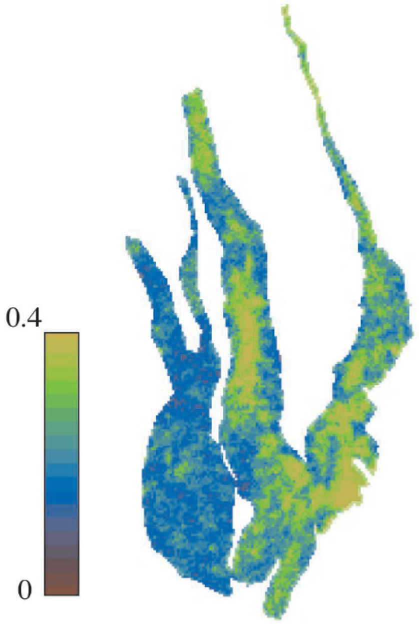


Figure 1. Vertically-averaged porosity in a Nigerian reservoir, obtained from seismic information and hard conditioning data

The alternative to sequential simulation is simulation by superposition of a kriging field plus a spatially correlated perturbation based on the orthogonality between kriging estimates and kriging errors (Journel, 1974). The steps are as follows:

- Perform kriging with linear constraints as indicated in the previous section.
- Generate an unconditional simulation of Z and Y. Since it is unconditional, first Z is simulated, then expression equation (1) is applied to the simulation to obtain the simulated Y values.

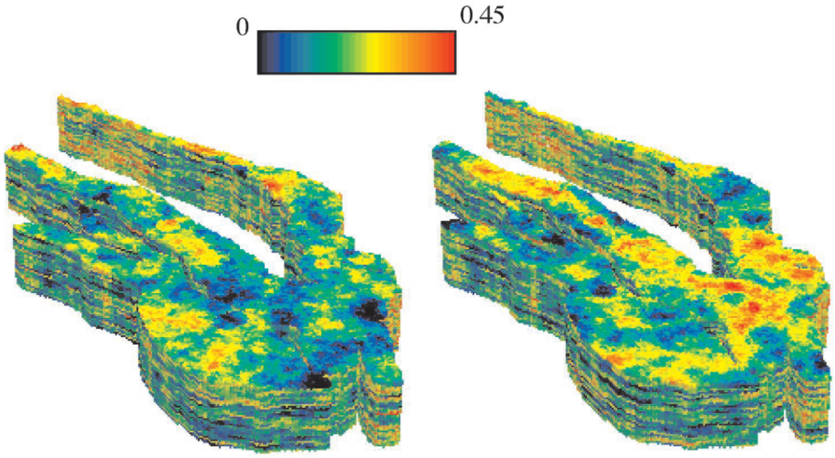


Figure 2. Two realizations of porosity in a Nigerian reservoir, discretized with a mesh of 200 by 400 blocks horizontally and 20 layers vertically

– Sample the unconditional simulation Z at the z conditioning data locations, plus the y unconditional values, and perform kriging with linear constraints as indicated in the previous section with this new set of data.
– Subtract the last kriging map from the unconditional simulation and add the differences to the first kriging map.

Since the linear constraints are honored in the three first steps of the algorithm, the final realization is a conditional realization in which the linear constraints are exactly reproduced.

4 Example

Starting from a geophysical survey and some hard conditioning data, the map in Figure 1 was generated as a representation of the vertically-averaged porosity in a Nigerian reservoir.

Using the map in Figure 1 a set of 3D realizations were generated for the reservoir with 20 layers and discretized with a mesh of 200 by 400 horizontally. Each realization is the result of a stochastic simulation of porosity with a large horizontal to vertical anisotropic ratio, conditioned to well data, and conditioned to the vertical averages in Figure 1. In Figure 2, two such realizations are displayed.

When taking the vertical average of the porosities in the realizations in Figure 2, the resulting average porosity is shown in Figure 3.

Comparison of Figure 3 and 1 indicates that the constraints have been exactly honored in both realizations. Furthermore, a quantitative evaluation of the conditioning yields a coefficient of correlation of 1.0 between the imposed constraints and the vertically-averaged simulations.

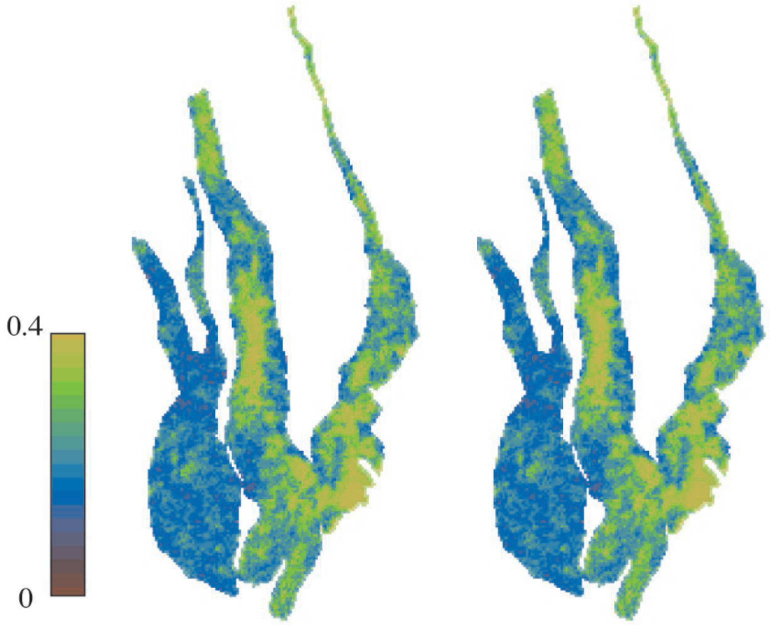


Figure 3. Vertical average of the porosity values in Figure 2

5 Conclusions

Although imposing linear constraints on kriging or simulations appears as a simple task in theory, exact reproduction of the constraints requires the re-evaluation of the approximations that are embedded in all estimation and simulation algorithms, especially the concept of search neighborhood.

Once these issues are addressed, the algorithm will reproduce, exactly, these constraints.

Acknowledgements

Total is gratefully acknowledged for allowing the use of its proprietary GeoChip code and the data from a Nigerian reservoir to illustrate this paper.

References

Deutsch, C. V. (1993). Conditioning reservoir models to well test information. In Soares, A. (Ed.), *Geostatistics Tróia '92, volume 1*, pp. 505–518. Kluwer Academic Publishers.

Gómez-Hernández, J. J., & Cassiraga, E. F. (2000). Sequential conditional simulation with linear constraints. In Kleingeld, W. J., & Krige, D. G. (Eds.), *Geostats 2000 Cape Town*, pp. in CD–ROM. Document Transformation Technologies, South Africa.

This page intentionally blank

Journel, A. G. (1974). Geostatistics for conditional simulation of ore bodies. *Economic Geology*, *69*(5), 673–687.

Journel, A. G. (1999). Conditioning geostatistical operations to nonlinear volume averages. *Math. Geology*, *31*(8), 931–953.

Srinivasan, S., & Journel, A. G. (1998). Direct simulation of permeability fields conditioned to well test data. In *Report 11*, Vol. 1. Stanford Center for Reservoir Forecasting, Stanford University.

A DIMENSION-REDUCTION APPROACH FOR SPECTRAL TEMPERING USING EMPIRICAL ORTHOGONAL FUNCTIONS

ALEXANDRE PINTORE and CHRIS C. HOLMES
Department of Statistics, University of Oxford, 1 South Parks Rd, Oxford, OX1 3TG, U.K.

Abstract. Recently, the authors introduced a tempering framework for constructing non-stationary analytical spatial covariance functions. However, a problem when using tempering of the spectrum obtained from the Karhunen-Loève expansion of a covariance matrix lies in the fact that the computational cost of the estimation procedure is of order $O(n^3)$, where n is the size of the data set. Prediction is also an important issue since it is again of order $O(n^3)$. We show that a method for approximating the eigenvectors of the Karhunen-Loève expansion at any location allows us to considerably reduce the computation required. This is achieved by selecting $m << n$ data points from the original data set and estimating the process using these points. Prediction is then carried out using an approximation to the full design matrix of the equivalent basis functions representation of our model. The resulting computational cost of our method is of order $O(m^2n)$. We report results on a Swiss data set and show that setting $m << n$ induces good accuracy in the solution.

1 Introduction

Modelling the second-order structure of Gaussian Processes is an important task in Geostatistics. Most current methods assume stationarity, which is equivalent to stating that the statistical association between two points is solely a function of the vector distance between them. However, the assumption of stationarity rarely holds in practise, and more often than not is made for mathematical convenience.

Recently, Pintore and Holmes (2004) proposed a new method for building non-stationary analytical covariance functions, which relies on the simple idea that one can easily build valid and interpretable non-stationary models by allowing the spectrum, taken in a wide sense, of some stationary process to evolve over space. They consider two representations of a covariance function, namely the Fourier representation and the Karhunen-Loève expansion. It is shown that when using the Karhunen-Loève expansion, a simple and natural way to allow for the spectrum (i.e. the eigenvalues) to evolve over space is by tempering it using a latent spatial process $\eta(\cdot)$ defined over the whole field. That is, by heating or cooling the spectrum

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 1007–1015.

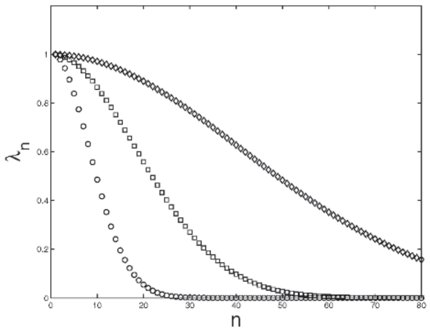


Figure 1. Plots of the normalised eigenvalues for the one-dimensional example for $\eta = 1$ (squares), $\eta = 5$ (circles) and $\eta = 0.2$ (diamonds), Section 1.

at each location, one is able to control the amount of smoothness induced by the model. An example is given in Figure 1 where the non-parametric Karhunen-Loève spectrum $\lambda_i,\ i = 1, ..., n$ is taken successively to the power $\eta = 0.2$, $\eta = 1$ and $\eta = 5$.

The authors are then able to show that in the finite dimensional case, the resulting covariance matrix has an analytical form given by

$$C_{NS}(s_i, s_j) = \sum_{k=1}^{n} \lambda_k^{\eta(s_i)/2} \lambda_k^{\eta(s_j)/2} \phi_k(s_i)\phi_k(s_j),\ i,\ j = 1, ..., n, \tag{1}$$

where $\lambda_k,\ k = 1, ..., n$ and $\phi_k(\cdot),\ k = 1, ..., n$ are respectively the eigenvalues and eigenvectors of the eigendecomposition of the "best" fit initial stationary matrix. Thus they are able to obtain interpretable non-stationary models using empirical orthogonal functions when only one measurement is available at each location.

With respect to the modelling of $\eta(\cdot)$, the authors suggest using a Bayesian regression on $\log \eta(\cdot)$ to ensure that $\eta(\cdot)$ remains strictly positive across the field, so that

$$\log\ \eta(s) = \beta_0 + s\beta_1 + \sum_{j=1}^{k} \psi(s, u_j)\beta_{j+2} \tag{2}$$

where β_0, β_1 capture linear trends in $\log \eta(s)$ and the regression splines $\psi(s, u_j)$ with knots u_j allow for spatial variation in $\eta(s)$. The following prior is assumed on the parameters,

$$\beta = (\beta_0, ..., \beta_{k+2})' \sim N(\mathbf{0}, b^2\mathbf{I}), \tag{3}$$

where b^2 is fixed to some reasonable value. Note however that care must be taken in order to avoid over-smoothing.

Note that in this paper we focus on modelling non-stationarity in the stochastic component of the process, which corresponds to "small-scale" variations in the structure of the spatial data. Our method could however be accommodated to consider the modelling of a deterministic trend component, corresponding to the "large-scale" variations in the process. More detail on the decomposition of a process into "large" and "small" scales, as well as on the way to account for a non-constant mean within the spectral tempering framework are given in Pintore and Holmes (2004).

1.1 PREDICTION

When predicting using (1), one needs to evaluate the covariance function between any two points in the space and thus extend the eigenvectors $\phi_i,\ i = 1, ..., n$ to eigenfunctions $\tilde{\phi}_i(\cdot),\ i = 1, ..., n$ defined over the whole space. This issue is discussed in detail in Pintore and Holmes (2004). Briefly, from the theory of the numerical treatment of integral equations (Baker, 1977), two classes of methods can be used to extend $\phi_i,\ i = 1, ..., n$, the integration formulae and expansion methods respectively. The expansion method was used in Obled and Creutin (1986) and consists in approximating $\tilde{\phi}_i(\cdot),\ i = 1, ..., n$ by a linear combination of linearly independent *a priori* chosen functions. These methods are computationally expensive however and the choice of the a priori functions is often not straightforward.

The integration formulae methodology on the other hand leads to the following formula for extending the eigenvectors,

$$\tilde{\phi}_i(s) = (1/\lambda_i) \sum_{j=1}^{n} C(s, s_j)\phi_i(s_j),\ i = 1, ..., n \tag{4}$$

where $C(\cdot, \cdot)$ is the "best" fit stationary matrix. The extension given by (4) is called the *Nyström extension* of the eigenvectors $\phi_i,\ i = 1, ..., n$. This method was used by Williams and Seeger (2001). Details on the derivation of (4) are given in Pintore and Holmes (2004).

We give examples of the predictions obtained in the case of unequally spaced data in Figures 2 and 3. The figures were drawn using $n = 50$ randomly chosen data points in [-5,5] and for four different types of covariance functions: the Matèrn with $\nu = 0.25$, $\nu = 0.5$ (i.e. Exponential), $\nu = 1$ and $\nu \to \infty$ (i.e. Gaussian). We expand the first and fourth eigenvectors in each Figure respectively. Note that we focus on the one-dimensional case for illustration purposes only.

It is then clear that one is able to compute the covariance between any two points using $\tilde{\phi}_i(\cdot),\ i = 1, ..., n$, so that, using kriging (Cressie, 1993), prediction is straightforward.

1.2 ESTIMATION

Pintore and Holmes (2004) estimate the parameters of the latent spatial process $\eta(\cdot)$ modelled as (2), using the REstricted Maximum-Likelihood (REML) described for instance in McCulloch and Searle (2001), that is,

$$l_m(C_{NS}, \mathbf{Z}) \propto -\frac{1}{2} \log |C_{NS}| - n/2 \log(\mathbf{Z}' C_{NS}^{-1} \mathbf{Z}). \tag{5}$$

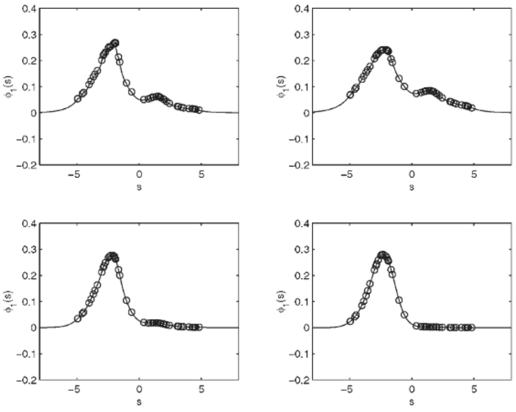


Figure 2. Numerical estimates of the 1st eigenfunction at 1000 equally spaced points in $[-8, 8]$ shown by the solid line, using 50 unequally spaced points (circles) in $[-5, 5]$ for the Matèrn 1 (top left), the Exponential (top right), the Matèrn 2 (bottom left) and the Gaussian (bottom right) covariance function, Section 1.1.

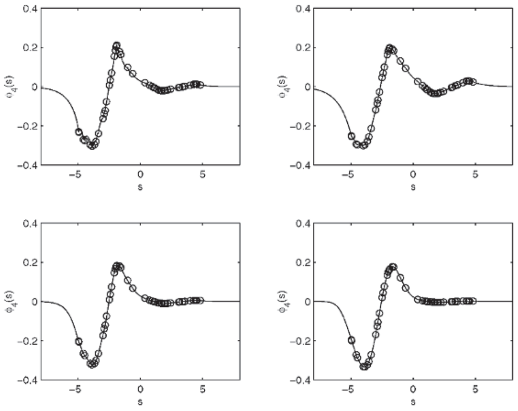


Figure 3. Numerical estimates of the 4th eigenfunction at 1000 equally spaced points in $[-8, 8]$ shown by the solid line, using 50 unequally spaced points (circles) for the Matèrn 1 (top left), the Exponential (top right), the Matèrn 2 (bottom left) and the Gaussian (bottom right) covariance function, Section 1.1.

To maximise the marginal log-likelihood we can use any practical method of optimisation, see Fletcher (1987) for examples. Pintore and Holmes (2004) choose to use a Nelder-Mead optimisation algorithm.

1.3 COMPUTATIONAL ISSUES

One of the main drawbacks of the spectral tempering method lies in the computational cost of the estimation procedure. Indeed, when maximising (5), each step of the optimisation procedure requires the inversion of the $n \times n$ matrix C_{NS}, which is of order $O(n^3)$. This leads to an important computational burden on the method when n is large.

This issue is also of concern in the stationary case when one seeks to estimate the parameters of the stationary covariance matrix using REML. Each step of the optimisation procedure again requires to invert an $n \times n$ matrix, the only difference with the non-stationary case being that one usually has fewer parameters to estimate so that fewer steps are required to maximise (5).

Note finally that prediction using kriging also requires the inversion of a $n \times n$ matrix in both cases which adds to the computational burden of the method. We consider these issues in the next section.

2 A rank-reduced procedure

2.1 APPROXIMATIONS IN THE STATIONARY CASE

With respect to the computational issues mentioned in the previous section, it is insightful to consider rank-reduced methods in the stationary case first. Two methods have been proposed which are reviewed in Williams and Seeger (2001).

- The first method consists in selecting only the first $m << n$ eigenvectors in order to model the stationary covariance matrix and to make predictions. The method was briefly discussed in Cohen and Jones (1969) and its overall computational cost is of order $O(m\ n^2)$. One of its main features lies in the fact that it uses information from all data points.
- The second method is more recent and makes use of the fact that one is able to extend the eigenvectors according to (4). The idea is to first *randomly* select $m << n$ points from the initial data set and fit a stationary model to these m points using only $p < m$ eigenvectors. Then the second step consists in approximating the full stationary covariance matrix using (4) to extend the p eigenvectors. Using the *Woodbury formula* (Press et al, 1992) to invert the resulting approximated covariance matrix leads to a method of order $O(m^2 n)$. Williams and Seeger (2001) show that this method gives accurate results for $m << n$ when averaged over a large number of runs for different m points randomly chosen on each run. However, while the average is accurate, for a single run the method appears to be quite sensitive to the choice of the m points.

2.2 APPROXIMATION IN THE NON-STATIONARY CASE

When considering the non-stationary case, care is needed. Indeed, recalling that spectral tempering induces non-stationary by spatially adapting the relative weight of high frequencies with respect to low frequencies, it is clear that the first method described in the previous subsection is not applicable. For if we select only the first p eigenvectors, that is the low frequency components of the process, then spatial adaptiveness via tempering of the spectrum loses much of its appeal since all the high frequency components are being ignored. As a consequence, we will use a method similar in spirit to the second method described in the stationary case but with important differences.

First, when choosing the initial m locations among the data points, we favour using the following procedure rather than selecting them randomly:

1. Obtain m locations $c_i, i = 1, ..., m$ in the field using k-means clustering on the full data set.
2. Because the initial locations must be data points, we then choose, for each cluster i, the data point which is closest to c_i.

We believe this procedure to be more efficient in order to obtain an initial set of m data points well distributed among the data set.

Once this set of m data points is selected, then we perform the spectral tempering method on this set in the usual way. The estimation procedure is thus now of order $O(m^3)$ rather than $O(n^3)$. This represents an important computational gain when n is large.

With respect to the prediction, an important issue lies in the fact that unlike in the stationary case, we are unable to use the Woodbury formula directly here. However, we show that our non-stationary modelling framework has features which allow us to still be able to predict in $O(nm^2)$.

2.2.1 *Prediction*

In Pintore and Holmes (2004), it is shown that spectral tempering using Empirical Orthogonal Functions has an equivalent basis functions representation where the variance of the random coefficients evolves over space. The authors refer to this model as a *variance-varying coefficients model.* Thus, we are able to write our model in the form

$$z(s_i) = \sum_{k=1}^{n} \beta_k(s_i)\phi_k(s_i), \;\; i = 1, ..., n \tag{6}$$

where $z(s_i), \;\; i = 1, ..., n$ is the (random) value of the spatial stochastic process under study at location $s_i, \;\; i = 1, ..., n$ and $\beta_k, k = 1, ..., n$ are independent mean-zero Gaussian processes such that for all $k = 1, ..., n$,

$$Cov(\beta_k(s_i), \beta_k(s_j)) = \lambda_k^{(\eta(s_i)+\eta(s_j))/2}, \;\; i, j = 1, ..., n \tag{7}$$

Now, the model given by (6)-(7) can be rewritten in the form

$$z(s_i) = \sum_{k=1}^{n} \beta_k \phi_k^{'}(s_i), \;\; i = 1, ..., n \tag{8}$$

where for all $k = 1, ..., n$,

$$\begin{aligned} \beta_k &\sim_{i.i.d.} N(0, 1) & (9) \\ \phi_k^{'}(\cdot) &= \lambda_k^{\eta(\cdot)/2} \phi_k(\cdot). & (10) \end{aligned}$$

that is, as a regression model with fixed random coefficients and spatially adaptive kernels. This in turn allows us to make predictions of order $O(nm^2)$ using the following procedure.

1. Consider the model (8) obtained from the m selected data points.
2. We extend it to the full data set by constructing the approximated spatially adaptive design matrix $\tilde{\Phi}^{'}$ of size $n \times n$ for all data points. This is done by considering (10) and using (4) to extend the eigenvectors and (2) to predict the $\eta(\cdot)$ process at all locations. This operation is of order $O(nm^2)$.
3. The p locations at which we wish to predict can be represented as

$$\mathbf{Z}_{pred} = \Phi_{pred}^{'} \beta + \epsilon_{pred} \tag{11}$$

where $\mathbf{Z}_{pred} = (z_{pred}^1, ..., z_{pred}^p)^{'}$, $\Phi_{pred}^{'} = \{\tilde{\phi}_i^{'}(s_{pred}^j)\}_{i=1,...,m, j=1,...,p}$ is the approximated design matrix for the points at which we wish to predict, and

$$\epsilon_{\mathbf{pred}} \sim N(\mathbf{0}, \sigma^2 I_p). \tag{12}$$

From the theory of Bayesian linear regression, we obtain that

$$p(Z_{pred}|z(s_i), i = 1, ..., n) \sim N(\Phi_{pred}^{'} \mathbf{m}^*, \sigma^2(I_p + (\Phi_{pred}^{'}) V^* (\Phi_{pred}^{'})^T)) \tag{13}$$

where

$$\begin{aligned} \mathbf{m}^* &= (I_m + (\tilde{\Phi}^{'})^T (\tilde{\Phi}^{'}))^{-1} (\tilde{\Phi}^{'}) (y(s_1), ..., y(s_n))^T & (14) \\ V^* &= (I_m + (\tilde{\Phi}^{'})^T (\tilde{\Phi}^{'}))^{-1} & (15) \end{aligned}$$

We compare $E(Z_{pred}|y(s_i), i = 1, ..., n)$ with the kriging predictors (using all eigenvectors or only p) in the next subsection. The kriging predictors are obtained as described for the stationary case (i.e. approximating the full covariance matrix for the data) in the previous section and their derivation are of order $O(n^3)$ and $O(pn^2)$ respectively. Our predicting method is of order $O(m^3)$ since $(I_m + (\tilde{\Phi}^{'})^T (\tilde{\Phi}^{'}))$ is of size $m \times m$.

Thus, the computational cost of prediction is of order $O(m^3 + m^2 n) = O(m^2 n)$. Hence, since estimation is of order $O(m^3)$, the total cost of our method is $O(m^3 + nm^2) = O(m^2 n)$ We present some experimental results in the next subsection.

2.3 RESULTS

We now consider a two-dimensional data set, given in Dubois (1998). This data set was initially used to compare different spatial interpolation methods and contained $n = 467$ rainfall measurements made in Switzerland on the 8th of May 1986. The

Table 1. Performance (i.e. Coef. of Deter. R^2) of the GP, NSGP, NSGPK and NSGPBF models on the Swiss data set for different values of m, Section 2.3.

m	30	50	75	367
GP	0.8872	0.9075	0.9138	0.9247
NSGP	0.9029	0.9303	0.9320	0.9350
NSGPK	0.8996	0.9214	0.9279	0.9297
NSGPBF	0.8983	0.9248	0.9301	0.9324

focus of the study was on rainfall values since it appeared in a previous study that it was one of the main parameters defining the radioactive deposition in the case of the Chernobyl disaster. We refer the reader to Dubois (1998) for a full description of the data set.

Here, we take our data set to be $n = 367$ randomly extracted measurements. Then, for different values of m we build the following models,

1. the stationary model (GP)
2. the non-stationary model built using 5 splines (NSGP)
3. the non-stationary model built using 5 splines and for which prediction is made using only the first 25 eigenvectors (NSGPK)
4. the non-stationary model built using 5 splines and for which prediction is made using the basis functions approach described previously (NSGPBF)

We take the Gaussian covariance function for each model. We compare the models through their computed coefficients of determination from the predictions made at the $n = 100$ other data locations. Recall that the coefficient of determination describes the amount of variability in the data is explained by the model and is given by

$$R^2 = 1 - SSRes/SST, \tag{16}$$

where $SSRes$ is the residual sum of squares and $SSTotal$ the total sum of squares in the data to be predicted. The results are given in Table 1.

It can be seen that the results for small m are close to those obtained with the full data. Indeed, using only 30 data points, we account for around 90% of the variance compared with 92% for the full stationary model. This seems to suggest that our method is able to use all the information in the data. Moreover, the results using the basis functions representation are very close to the ones obtained using kriging which suggests it is a good approximation. More work and research will however have to be done on this subject. As suggested by an anonymous referee, a possible direction for future work could for instance involve comparing the mean square of the normalised (by the kriging standard deviation) residuals obtained for each model in order to get some measure of precision. Note that some of the important properties of the NSGP model are described in more detail in Pintore and Holmes (2004).

This page intentionally blank

3 Conclusions

In this paper, we have described a rank-reduced method for spectral tempering with Empirical Orthogonal Functions, which is of order $O(nm^2)$ where $m << n$, that is of the same order of rank-reduced methods in the stationary case. Prediction is carried out using an equivalent basis functions representation to the spectral tempering model. Although the primary results look promising, more experiments will have to be done in order to evaluate the efficiency of this method more accurately.

References

Baker, C. T. H., *The numerical treatment of integral equations*, Oxford: Clarendon Press, 1977.

Cohen, A. and Jones, R. H., *Regression on a random field*, J. Amer. Statist. Assoc., vol. 64, no. 328, 1969, p. 1172-1182.

Cressie, N., *Statistics for Spatial Data*, John Wiley, New York, 2nd, 1993.

Obled, C. and Creutin, J.D., *Some developments in the use of Empirical Orthogonal Functions for mapping meteorological fields*, J. of Clim. and Appl. Meteor., vol. 25, no. 9, 1986, p. 1189-1204.

Dubois, G., *Spatial Interpolation Comparison 97: Forward and Introduction*, J. of Geographic Inf. and Dec. Analy., vol. 2, 1998, p. 1-11.

Fletcher, R., *Practical Methods of Optimization*, John Wiley & Sons, 2nd, 1987.

McCulloch, C. E. and Searle, S. R., *Generalized Linear and Mixed Models*, New-York, Chichester: Wiley, 2001.

Pintore, A. and Holmes, C. C., 2004, *A New Framework for Constructing Non-stationary Geostatistical Covariance Functions*, Technical Report, Department of Statistics, University of Oxford.

Press, W. H., Teulosky, S. A., Vetterling, W. T. and Flannery, B. P., *Numerical Recipes in C*, Cambridge University Press, 2nd edition, 1992.

Williams, C. K. I. and Seeger, M., *Using the Nyström method to speed up kernel machines*, Advances in Neural Information Processing Systems, MIT Press, 2001.

A NEW MODEL FOR INCORPORATING SPATIAL ASSOCIATION AND SINGULARITY IN INTERPOLATION OF EXPLORATORY DATA

QIUMING CHENG
Department of Earth and Space Science and Engineering, Department of Geography, York University, Toronto,4700 Keele Street, Ont. M3J 1P3, Canada, E-mail: qiuming@yorku.ca
The State Key Laboratory of Lithosphere Evolution and Mineral Resources, China University of Geosciences, China,
E-mail:qiuming@cug.edu.cn

Abstract. Spatial association indices (autocorrelation, covariance, and variogram) have been largely used to characterize the local structure of surfaces and for data interpolation in kriging. Singularity is another index quantifying the scaling invariance property of measures from a multifractal point of view. Spatial association and scaling invariance characterize the local structure properties of surfaces from different aspects. Both must be taken into account in data interpolation and surface modeling. Kriging as one mapping technique is based on the spatial association of neighbourhood values through semivariogram. Recent study of multifractal modeling has shown that the local singularity exponent can quantify the local scaling invariance property characterizing the concave/convex properties of the neighbourhood values. The method proposed in this paper can incorporate both the singularity and spatial association in data interpolation. The ordinary kriging only becomes the special situation of the new method when it deals with nonsingular measures. It has been shown by a case study of the geochemical distributions of As, Cu, Pb, and Ag in lake sediment samples from southwestern Nova Scotia, Canada, that combining spatial association with singularity can improve the interpolation results significantly, especially for observed values with significant singularity.

1 Introduction

Geostatistics has long been applied in geosciences for data estimation and simulation. It involves semivariogram as the basic function to measure the spatial association and spatial variability of data. Semivariogram as a function of distance between locations can measure the spatial auto-correlation between values at locations separated by a distance. Various types of functions or models can be fitted to semivariogram and then used for assigning weights for weighted averaging in kriging. The main purpose of using kriging is for data interpolation, a process of assigning values for those locations where there are no observed values available from their neighbourhood observed values. Semivariogram has also been used for characterizing the structural property of landscape (Journel & Huijbregts, 1978). The method developed by Cheng (1999a)

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 1017–1025.

integrates both spatial association and local singularity, therefore, can enhance and retain the local structure properties when applied to 1-D data interpolation. It combines the semivariogram quantifying the spatial association and the singularity index characterizing the local structure of data. This idea was extended to 2-D situation so that surfaces can be created from interpolating 2-D point data (Cheng, 2000, 2001). A more general mathematical model is introduced in this paper and it is demonstrated by a case study of mapping geochemical concentration values from 1948 lake sediment samples from southwestern Nova Scotia, Canada.

2 Spatial Associations Vs. Singularity

2.1 SPATIAL ASSOCIATION

Spatial association represents a type of statistical dependency of values at separate locations. If the value at a location is considered as a realization of a so-called regionalized random variable, the spatial association or variability can be measured by means of semivariogram as

$$2\gamma(\mathbf{x},\mathbf{h}) = E\{[Z(\mathbf{x}) - Z(\mathbf{x}+\mathbf{h})]^2\} \qquad \textbf{(1)}$$

Where $\gamma(\mathbf{x}, \mathbf{h})$ is a function of vector distance **h** separating locations **x** and **x** + **h.** The semivariogram measures the symmetrical variability between Z(**x**) and Z(**x** + **h**). Under an assumption of the second-order stationarity, the semivariogram (1) becomes the function of h independent of location x. This strong assumption of the regionalized random variable is generally required in kriging. The function (1) has been commonly used for structural analysis and interpolation in geostatistics (Journel & Huijbregts, 1978). It has also been applied for texture analysis in image processing (Atkinson & Lewis, 2000; Herzfeld, 1993; Herzfeld & Higginson, 1996).

2.2 SINGULARITY

The singularity in the multifractal context characterizes how the statistical behaviour varies as measuring scale changes. For example, in some locations the mean values calculated from the neighborhood values might be independent of the size of the vicinity within which the values are averaged. In other cases the mean value might proportionally depend on the size of the vicinity. We call the former case nonsingular location and the latter singular location. Singularity property has been commonly observed in geochemical and geophysical quantities (Cheng, Agterberg and Ballantyne, 1994; Cheng, 1997, 1999a, 2000). Taking the notation of multifractal model, the singularity index $\alpha(\mathbf{x})$ is related to the measure defined in a small vicinity around location **x** of linear size ε, $\Omega(\mathbf{x}, \varepsilon)$, as

$$\mu(\varepsilon) = c\varepsilon^{\alpha(\mathbf{x})} \qquad \textbf{(2)}$$

Where c is constant. In case of geochemical data, $\mu(\varepsilon)$ can be defined as the amount of metal in an area of size ε. For convenience without loss of generality, we will introduce a density function $\rho(\varepsilon)$ as

$$\rho(\varepsilon) = \mu(\varepsilon)/\varepsilon^{E} = c\varepsilon^{\alpha(\mathbf{x})-E} \qquad \textbf{(3)}$$

Where E is the dimension of the vicinity $\Omega(\mathbf{x}, \varepsilon)$ (E=2 for 2-D geochemical map). For 2-D problems, the vicinity can be chosen as circular or square in shapes. The value of the singularity α ranges from $\alpha_{\min}$ to $\alpha_{\max}$. The index α can be estimated by least square method to fit a straight line to a set of values $\mu(\varepsilon)$ against ε on a log-log paper. This can be done directly from the original point sample data so that it is not affected by the smoothing of interpolation. The value can be taken as the slope of the straight line. The error involved in the estimation can be calculated from least square fitting. The singularity index estimated from equation (3) has the following properties (Cheng, 1999a):

1. $\alpha = E$, iff $\rho(\varepsilon)$ = constant, independent of vicinity size ε.
2. $\alpha < E$, iff $\rho(\varepsilon) \propto \varepsilon^{\alpha-E}$ is a decreasing function of ε, which normally implies the "convex" property of $\mu(\varepsilon)$ at the location **x**.
3. $\alpha > E$, iff $\rho(\varepsilon) \propto \varepsilon^{\alpha-E}$ is an increasing function of ε, which indicates the "concave" property of $\mu(\varepsilon)$ at the location **x**.

The cases (2) and (3) correspond to singular situations in which the density function $\rho(\varepsilon) \to\infty$ or $\rho(\varepsilon) \to 0$ as $\varepsilon \to 0$. In the case of $\rho(\varepsilon) \to\infty$, it implies that within a small area (small ε) there is an anomalously high density of element concentration. These singular locations are often associated with anomalies in geochemical exploration caused by mineralization. The anomalous area may have enriched concentration many times higher than the background values. Therefore, the index α can be used as measure characterizing the structural property of measure $\mu(\varepsilon)$. This index, like other indexes, measures the local structural property at certain scale which is determined by the scaling range used for the estimation of the value index. It needs a good coverage of points in order to accurately estimate the singularity index in a small scale. It has been used for texture analysis to remote sensing Landsat TM image (Cheng, 1997, 1999b), in multifractal interpolation of geochemical concentration values for mineral exploration (Cheng, 1999a, 2000) and in well log curve reconstruction (Li and Cheng, 2001).

2.3 DISTRIBUTION OF SINGULARITY INDEX

The singularity index usually has finite values around E. For a conservative multifractal measure, the dimension of the set with $\alpha = E$ is close to E (box-counting dimension) which means that the areas on a geochemical map with nonsingular values occupy most part of the map. The dimensions of the other areas with $\alpha(\mathbf{x}) \neq E$ are given by the fractal spectrum function $f(\alpha) < E$ (Cheng, 1999a). This implies that the areas with singular values (anomalies) are relatively small in comparison with the areas with non-singular values (background values). From a statistical point of view, the majority of values on the geochemical map where $\alpha \approx 2$ follow either normal or lognormal distribution whereas the extreme values on the map with singularity $\alpha(\mathbf{x}) \neq E$ may follow fractal distributions. To remove the samples with extreme values from the inputs for kriging has been the common practice in data interpolation. However, for exploration purpose the removal of the singular samples will smooth off local variability that may carry valuable information for anomaly identification in mineral exploration. The ability of dealing with singular values must be the qualification of quantitative methods for handling exploratory dataset in mineral exploration. Most ordinary statistics requiring assumption of normal (lognormal) distribution of values may not be effective in dealing exploratory dataset with extreme value distribution. Multifractal modeling techniques have been demonstrated as possible to solve the above problems.

3 General Model Incorperating Both Spatial Association And Singularity

Scaling property has been commonly observed in various types of patterns in geosciences. Use of scaling property for prediction and estimation purposes has attracted tremendous attention. Statistical property derived at one scale may be used to estimate the property in another scale on the basis of the scaling property. Data interpolation including kriging is to estimate values at unknown locations and this type of process can be considered as down scaling process. How to apply scaling property in the process is obviously of general interests. The multifractal interpolation method developed by Cheng (1999a, 2000) for construction of curves (1-D problem) and surface (2-D) on the basis of point observations uses the scaling properties of geochemical data. It incorporates the local singularity as well as the spatial association of the data in the data interpolation. This paper introduces a general mathematical model of the method and discusses its advantages and disadvantages.

Relation (3) shows that the density function at a given location x follows a power-law relationship with the scale unit (box size ε). The exponent $\alpha(x)$ –E characterizes the local singularity of the function – how the function changes as the scale unit decreases. At the singular location, $\alpha(x) \neq E$, the density is dependent of the scale unit. In this case the constant c becomes a useful quantity independent of scale unit which can be considered as the measure of the density $\rho(\varepsilon)$ in the space of $\alpha(x)$ –E dimension. It is no longer singular value. The value c becomes the ordinary density value in non-singular locations; therefore, the quantity c instead of $\rho(\varepsilon)$ can be used to form the interpolation formulism.

To derive the new interpolation relation to incorporate both spatial association and singularity, let us arbitrarily choose vicinity $\Omega(\mathbf{x_0}, \varepsilon)$, a small area on 2-D geochemical map. For convenience without loss of generality, we will introduce a notation $Z(\boldsymbol{\varepsilon})$ to represent the average geochemical concentration value at location **x** within $\Omega(x, \varepsilon)$. $Z(\boldsymbol{\varepsilon})$ is a type of density measurement with common units of ppb, ppm or %. Substitute $\rho(\varepsilon)$ with $Z(\boldsymbol{\varepsilon})$ in (3) gives

$$Z(\varepsilon) = c(x)\varepsilon^{\alpha(x)-E} \quad \textbf{(4)}$$

Several values of $Z(\varepsilon)$ with variable ε can be used to estimate the constant quantity c(**x**) at location **x**. As discussed previously, the new density quantity c(**x**) becomes a non-singular quantity. We can establish a relation between the density quantity $c(x_0)$ at the center and their neighborhood values $c(x_i)$ as following

$$c(x_0) = \sum_{x_i \in \Omega(x_0, \varepsilon)} \lambda_i(\|x_i - x_0\|)\, c(x_i) \quad \textbf{(5)}$$

where c($\mathbf{x_0}$) and c($\mathbf{x_i}$) are density quantities estimated from (3) at locations $\mathbf{x_0}$ and $\mathbf{x_i}$, respectively, $\lambda_i(\|\mathbf{x_i}\text{-}\mathbf{x_0}\|)$ is the weighting factor to be determined, $\sum \lambda_i(\|x_i - x_0\|) = 1$. The value of λ can be estimated using inverse distance weighting or kriging methods. Since the estimation of weighting factor λ is well known process, there is no need to repeat the actual estimating process. Here we will derive the new interpolation model and to compare it with the ordinary model. For a chosen resolution ε_* ($\varepsilon_* \leq \varepsilon$) for generating the interpolating map, the average value of $Z(\varepsilon_*)$ within a vicinity of size ε_* (pixel on the interpolation map) can be replaced by the actual value $Z_{\mathbf{x}}$ observed at the center of the vicinity which can be related to the estimation c($\mathbf{x}$) $\varepsilon_*^{\alpha(\mathbf{x})-E}$. Substitute the observed values of geochemical density to the density quantity c-values yields

$$Z_{\mathbf{x_0}} \varepsilon_*^{E-\alpha(\mathbf{x_0})} = \sum_{\mathbf{x_i} \in \Omega(\mathbf{x_0}, \varepsilon)} \lambda_i(\|\mathbf{x_i} - \mathbf{x_0}\|) Z_{\mathbf{x_i}} \varepsilon_*^{E-\alpha(\mathbf{x_i})}$$

or

$$\begin{aligned} Z_{\mathbf{x_0}} &= \varepsilon_*^{\alpha(\mathbf{x_0})-E} \sum_{\mathbf{x_i} \in \Omega(\mathbf{x_0}, \varepsilon)} \lambda_i(\|\mathbf{x_i} - \mathbf{x_0}\|) Z_{\mathbf{x_i}} \varepsilon_*^{E-\alpha(\mathbf{x_i})} \\ &= \sum_{\mathbf{x_i} \in \Omega(\mathbf{x_0}, \varepsilon)} \lambda_i(\|\mathbf{x_i} - \mathbf{x_0}\|) \varepsilon_*^{\alpha(\mathbf{x_0})-\alpha(\mathbf{x_i})} Z_{\mathbf{x_i}} \end{aligned} \qquad \textbf{(6)}$$

Relation (6) is a general weighted average model that can be used to estimate the value ($Z_{\mathbf{x0}}$) at the center of $\Omega(x_0, \varepsilon)$ from the neighborhood values ($Z_{\mathbf{xi}}$) within $\Omega(x_0, \varepsilon)$. It has the following properties:

1. If the entire dataset does not show singularity, $\alpha \equiv E$, then (5) and (6) are identical and the same as the ordinary moving average function that has been used commonly in kriging and other data interpolation methods.
2. If all values in the entire vicinity show the same singularity strength, α = constant, then (6) becomes the same as the ordinary moving average function used in kriging and other methods.
3. If the neighborhood values are non-singular but not the value at the center, $\alpha(x_0) \neq E$, and $\alpha(x_i) = E$, then (6) is equivalent to the ordinary moving average function multiplied by a factor $\varepsilon^{\alpha(x_0)-E}$.

$$Z_{x_0} = \varepsilon^{\alpha(x_0)-E} \sum_{x_i \in \Omega(x_0, \varepsilon)} \lambda_i(\|x_i - x_0\|) Z_{x_i} \qquad \textbf{(7)}$$

The factor $\varepsilon^{\alpha(x_0)-E}$ modifies the ordinary average in such that if $\alpha(x_0) < E$, then the new result is increased by a factor $\varepsilon^{\alpha(x_0)-E}$ given small ε, whereas if $\alpha(x_0) > E$, then the new result is reduced by a factor $\varepsilon^{\alpha(x_0)-E}$. This modification is reasonable because $\alpha < E$ and $\alpha > E$ correspond to convex and concave properties of surface $Z_{\mathbf{x}}$ around the location **x**, respectively. The relation (7) was introduced in the author's previous publication (Cheng, 2000, 2001). The model not only involves the spatial association reflected in the calculation of weight λ but also incorporates the singularity characterized by the singularity index α. It is obvious that the ordinary weighted average model (used by IDW and kriging) becomes the special case of the new method expressed in (6). The new model has, therefore, two obvious advantages: it not only improves the accuracy of the interpolated results but also retains the local structure of the interpolated map. The latter is essential for geochemical and geophysical data processing and pattern recognition. This will be demonstrated using geochemical concentration values of as from 1948 lake sediment samples from southwestern Nova Scotia, Canada.

4 Mapping Geochemical Values Of As From Lake Sediment Samples

Geochemical data from 1948 lake sediment samples have been analyzed using various statistical and multifractal techniques for detection of Au, U, Sn and W mineralzation associated alteration zones in the southwestern Nova Scotia, Canada (Xu & Cheng, 2001). The geology of the study area is illustrated in Fig. 1. The study area (≈4000 km^2) is mainly underlain by Cambro-Ordovicien low-middle grade metamorphosed sedimentary rocks and Devonian granitoid rocks. The South Mountain Batholith (SMB) is a complex of multi-phase granites covering nearly one-third of the entire study area. A number of Au, W, and Sn deposits have been found in the area. About 45 Au mineral deposits are shown as dots in Fig. 1. More detailed discussion of the geology and geological controlling features on the spatial distribution of Au deposits can be found in Xu & Cheng (2001). For demonstration purpose, the values of As from the lake sediment samples will be mapped both by the ordinary kriging and by the method introduced in this paper.

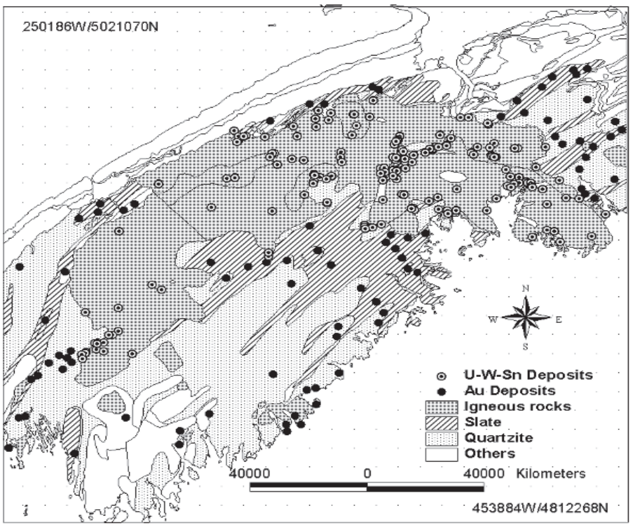


Figure 1. Simplified geology in southwestern Nova Scotia, Canada

Fig. 2 shows the map generated from 1948 As values by means of ordinary kriging with spherical model with a search distance 8 km and maximum interpolation point 16 (Xu & Cheng, 2001). Fig. 3 illustrates the distribution of α-values (< 2 as contours) estimated based on the distribution of As values from 1948 lake sediment samples with a maximum scaling range ε_{max} = 11km. The values

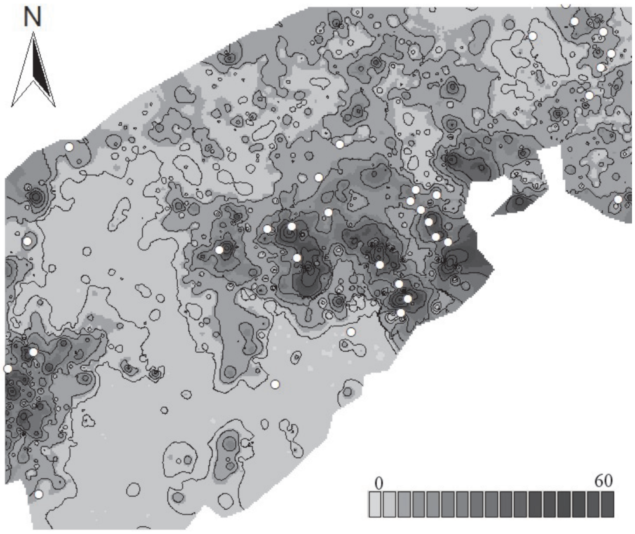


Figure 2. Kriging map of As. Detailed parameters can be found in the text. Black outlines represent the granitoid complex (SMB). Dots represent gold mineral deposits .

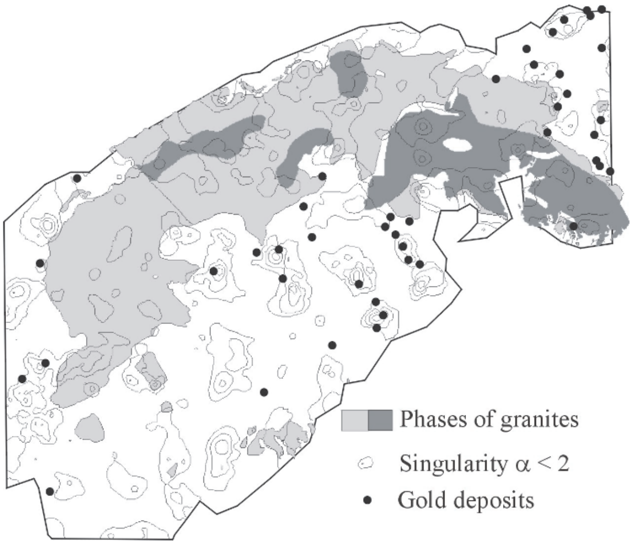


Figure 3. Estimated singularity values (α) for As. Contour lines represent area with singularity $\alpha < 2$. Dots present gold mineral deposits. The map is smaller than map in Figure 2 due to edge effect.

of the correlation coefficients related to the linear fitting in the estimation of α by plotting log $\mu(\varepsilon)$ against log ε for $\varepsilon = 2, \ldots, 11$ km are calculated and these values range from 0.97 to 1, implying significant linear relationships exist between log $\mu(\varepsilon)$ and log ε for all the locations. The results obtained using Eq. 7 are shown in Fig. 4. It can be seen that patterns with $\alpha < 2$ are mainly distributed either in the south of SMB as linear patterns with NW-SE orientation or aggregated around the contacts of SMB, specially in those places where exist faults or transition zones of different granitoid phases. Some of the clusters with low α-values show strong spatial correlation with the spatial locations of Au deposits. This should not be surprised since low α-value may indicate the area with the enrichment of geochemical values that might due to mineralization in this study area. The general patterns in Figs. 2 & 4 look similar, the ratio of these two maps, however, clearly show the differences of them. Similar as the patterns with $\alpha < 2$ in Fig. 3, the patterns with the ratio > 1 in Fig. 5 clearly highlight not only the linear geochemical anomalies in the south of SMB but also the areas around the SMB at where faults or transition of granitoid phases. The improvement shown in Fig. 5 may be significant when applied for anomaly interpretation that enhances the local structural properties of the mapped surface. The results obtained for other elements (Cu, Pb, and Zn) in the area also indicate that the areas with singular geochemical values ($\alpha < 2$) are favorable for Au mineralization in the study area (results not shown here).

5 Conclusions and Discussions

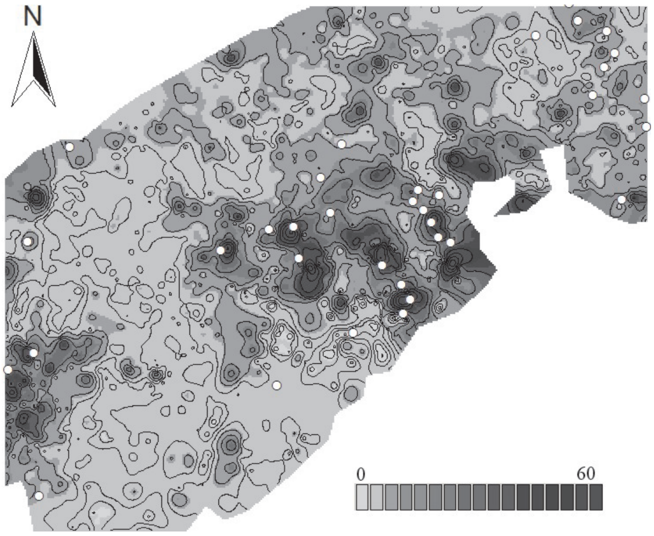


Figure 4. Results obtained using the multifractal interpolation method for As.

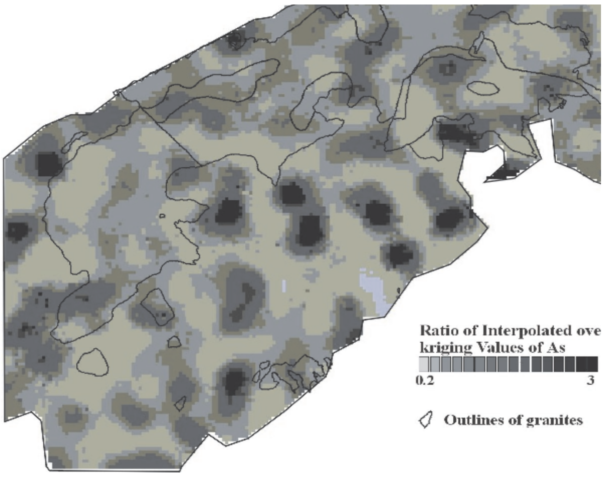


Figure 5. Ratio of kriging results over the results obtained using the multifractal interpolation method for As

The multifractal interpolation method proposed in this paper can be used for mapping purpose with the localized structural properties (multifractality) preserved. It has been demonstrated that this method is superior to the moving average techniques. The ordinary moving average methods can be considered as the special cases of the multifractal interpolation method when the interpolated data show nonsingular property. However, for most quantities in the exploratory geodatsets that show singularity, the ordinary moving average techniques including ordinary kriging are not applicable but the multifractal interpolation method can be used in order to retain the localized structural property. This paper has proposed a general

This page intentionally blank

function for cooperating both association and singularity. This may open a direction for studying how to improve the interpolation results by including both spatial association and singularity. Further study will be devoted to look into the anisotropy association and singularity properties and irregular instead of regular moving windows should be used in the estimation of singularity index and the moving average. Since the method involves the local singularity calculated according to the localized power-law of (4) using the original point sample data, it usually requires a dataset with good point coverage so that accurate singularity can be estimated. The error associated with the estimation will impact on the final interpolated result.

Acknowledgements

We would like to acknowledge the GeoDAS mining consortium for supporting this research in the Geomatics Research Lab at York University. This research has been partially supported by a Chinese 863 Research Program (2002AA135090). Thanks are due to the two anonymous reviewers for their constructive comments on critical reviewing the manuscript.

References

Atkinson, P.M., & Lewis, P., 2000. Geostatistical classification for remote sensing: an introduction, Computers&Geosciences, 26(4): 361-371.

Cheng, Q., 2001. A multifractal and geostatistical method for modeling geochemical map patterns and geochemical anomalies, Journal of Earth Sciences (in Chinese with English Abstract), 26(2): 161 – 166.

Cheng, Q., 2000. Interpolation by means of multiftractal, kriging and moving average techniques, in the Proceedings of GAC/MAC meeting GeoCanada2000, May 29 to June, 2, 2000, Calgary. http://www.gisworld.org/gac-gis/geo2000.htm

Cheng, Q., 1999b. Multifractality and spatial statistics. Computers&Geosciences, 25(9): 949-961.

Cheng, Q., 1999a. Multifractal interpolation: in S.J. Lippard, A. Naess and R. Sinding-Larsen (eds.) Proceedings of the Firth Annual Conference of the International Association for Mathematical Geology, Trondheim, Norway, v. 1, 245-250.

Cheng, Q., 1997. Multifractal modeling and lacunarity analysis. Math. Geology, 29(7): 919-932.

Cheng, Q., Agterberg, F.P., and Ballantyne, S.B., 1994. The separation of geochemical anomalies from background by fractal methods, Journal of Exploration Geochemistry, 51(2): 109-130.

Herzfeld, U.C., 1993. A method for seafloor classification using directional variogram, demonstrated for data from the western flank of the Mid-Atlantic Ridge. Math. Geology, 25(7): 901-924.

Herzfeld, U.C., & Higginson, C.A., 1996. Automated geostatistical seafloor classification-principles, parameters, feature vectors, and discrimination criteria. Computers&geosciences, 22(1): 35-52.

Journel, A.G., & Huijbregts, CH. J., 1978. Mining geostatistics, Academic Press, New York, 600p.

Li Q. and Cheng, Q. 2001. Fractal correction of well logging curves, Journal of China University of Geosciences, 12(3): 272-275.

Xu, Y. and Cheng, Q., 2001, A multifractal filter technique for geochemical data analysis from Nova Scotia, Canada, J. Geochemistry: Exploration, Analysis and Environment, 1(2): 1-12.

LOGNORMAL KRIGING: BIAS ADJUSTMENT AND KRIGING VARIANCES

NOEL CRESSIE and MARTINA PAVLICOVÁ
Department of Statistics, The Ohio State University, Columbus OH 43210-1247, USA

Abstract. Lognormality of spatial data occurs commonly enough for it to warrant continued study; contemporary statistical and computational methodologies can shed new light on the old problem of block kriging for lognormal processes. There are a number of proposals available for block kriging, many of them discussed in an unpublished, 43-page, Centre de Morphologie Mathematique "note" written by Georges Matheron in 1974. Loosely translated, the title of the note is, "The proportional effect and lognormality or: The return of the sea serpent". Our paper is meant to rein in the sea serpent, by comparing an optimal-prediction-based predictor with a permanence-approximation-based predictor, in the context of statistics for spatial lognormal data.

Key words: Empirical Bayes, geostatistics, MSPE, optimal spatial prediction

1 Introduction

Lognormal spatial data are common in mining and soil-science applications. Point kriging results in a map of the region of interest. However, mining and precision agriculture is carried out selectively and based on block averages of the process on the original scale. Finding spatial predictions of the blocks assuming a lognormal spatial process has a long history in geostatistics (Marechal, 1974; Matheron, 1974; Rendu, 1979; Journel, 1980; Dowd, 1982; Rivoirard, 1990; Roth, 1998). These papers cover the following topics: determine types of variogram models for lognormal data; decide whether to do inference on the original scale or the log scale; choose an optimality criterion for kriging; derive the kriging equations according to the optimality criterion; consider the cases of known or unknown mean (on the log scale); and consider whether knowing just the variogram (on the log scale) is enough to do kriging.

The purpose of this paper is to take a fresh look at two of the many possibilities for lognormal kriging, based on the following principles: the original scale is for optimality criteria (including unbiasedness) but the log scale is for linear statistical analysis; stationarity is needed for estimation of spatial dependence but it is not

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 1027–1036.

needed for spatial prediction (i.e., kriging); kriging is an empirical-Bayes methodology that requires efficient estimators of unknown parameters to be "plugged into" (simple) kriging equations.

A very influential piece of writing on lognormal kriging has been the unpublished 43 page "note" by Matheron (1974). Matheron's approach is to look at the problem from all sides, with many calculations drawn from Matheron (1962), but no definitive conclusions. His writing touches on all the geostatistical themes given above. At the time it was written, the statistical influences of linear models, efficient parameter estimation, and prediction theory were not yet felt in geostatistics. Moreover, computing power 30 years ago was a very small fraction of what it is today, making some forgotten block-kriging predictors feasible in today's computing environments. We concentrate on two predictors in this paper, and for each predictor we derive an analytical expression for its mean squared prediction error (i.e., kriging variance). Conditional simulation could produce equivalent results but without the easy interpretations that come with having analytical expressions.

2 Lognormal spatial process

Let the process $\{Z(\boldsymbol{s})\colon \boldsymbol{s} \in D\}$ denote a lognormal spatial process defined on a domain $D \subset \mathbb{R}^d$. That is,

$$Y(\boldsymbol{s}) \;\equiv\; \log Z(\boldsymbol{s})\,; \qquad \boldsymbol{s} \in D\,, \tag{1}$$

is a Gaussian process defined with first two moments, $\mu_Y(\boldsymbol{s}) \equiv E(Y(\boldsymbol{s}))$; $\boldsymbol{s} \in D$, and $C_Y(\boldsymbol{u},\boldsymbol{v}) \equiv \mathrm{cov}(Y(\boldsymbol{u}),Y(\boldsymbol{v}))$; $\boldsymbol{u},\boldsymbol{v} \in D$. Consequently, from (1), $Z(\boldsymbol{s}) = \exp\{Y(\boldsymbol{s})\} > 0$; $\boldsymbol{s} \in D$, and from Aitchison and Brown (1957),

$$\mu_Z(\boldsymbol{s}) \;\equiv\; E(Z(\boldsymbol{s})) = \exp\{\mu_Y(\boldsymbol{s}) + (1/2)C_Y(\boldsymbol{s},\boldsymbol{s})\}\,; \quad \boldsymbol{s} \in D\,, \tag{2}$$

$$C_Z(\boldsymbol{u},\boldsymbol{v}) \;\equiv\; \mathrm{cov}(Z(\boldsymbol{u}),Z(\boldsymbol{v})) = \mu_Z(\boldsymbol{u})\mu_Z(\boldsymbol{v})[\exp\{C_Y(\boldsymbol{u},\boldsymbol{v})\} - 1]\,. \tag{3}$$

From (2), $\mu_Z(\boldsymbol{s}) \geq \exp\{\mu_Y(\boldsymbol{s})\}$, giving rise to a potential source of bias when transforming back to the original scale. The presence of the mean terms as multipliers in (3) is sometimes called the proportional effect.

The spatial (lognormal) data are defined as the $(n \times 1)$ vector, $\boldsymbol{Z} \equiv (Z(\boldsymbol{s}_1),\ldots, Z(\boldsymbol{s}_n))'$, where $\{\boldsymbol{s}_1,\ldots,\boldsymbol{s}_n\}$ are known spatial locations. Then the transformed data, $\boldsymbol{Y} \equiv (Y(\boldsymbol{s}_1),\ldots,Y(\boldsymbol{s}_n))'$, are normally distributed and will be used to estimate unknown parameters in $\mu_Y(\boldsymbol{s})$ and $C_Y(\boldsymbol{u},\boldsymbol{v})$, as well as to predict an unknown value $Z(\boldsymbol{s}_0)$; $\boldsymbol{s}_0 \in D$. The prediction problem is sometimes called point kriging.

Geostatistics can be thought of as an empirical-Bayes methodology, where the "Bayes" part refers to putting a prior on the mean (e.g., Cressie, 1993, p. 171), and the "empirical" part refers to estimation of the fixed but unknown spatial covariance (or variogram) parameters. In Section 2.1, we consider the case of a singular prior (known mean), and in Section 2.2 we consider the case of a diffuse prior (generalized-least-squares estimation of the mean).

2.1 KNOWN MEAN (COVARIANCE FUNCTION ASSUMED KNOWN)

Assume $\mu_Y(\cdot)$ is known (and we always assume $C_Y(\cdot,\cdot)$ is known, although its unknown parameters are ultimately estimated). The minimum-mean-squared-prediction-error predictor $Y^*(\boldsymbol{s}_0)$, which in fact is the simple-kriging predictor, is

$$Y^*(\boldsymbol{s}_0) \;\equiv\; E(Y(\boldsymbol{s}_0)|\boldsymbol{Y}) = \mu_Y(\boldsymbol{s}_0) + \boldsymbol{c}_Y(\boldsymbol{s}_0)'\Sigma_Y^{-1}(\boldsymbol{Y} - \boldsymbol{\mu}_Y)\,, \tag{4}$$

where $\boldsymbol{c}_Y(\boldsymbol{s}_0) \equiv (C_Y(\boldsymbol{s}_0,\boldsymbol{s}_1),\ldots,C_Y(\boldsymbol{s}_0,\boldsymbol{s}_n))'$, $\Sigma_Y \equiv \mathrm{var}(\boldsymbol{Y})$, and $\boldsymbol{\mu}_Y \equiv (\mu_Y(\boldsymbol{s}_1), \ldots, \mu_Y(\boldsymbol{s}_n))'$. Notice that the simple-kriging variance is,

$$\mathrm{var}(Y(\boldsymbol{s}_0)|\boldsymbol{Y}) \;=\; C_Y(\boldsymbol{s}_0,\boldsymbol{s}_0) - \boldsymbol{c}_Y(\boldsymbol{s}_0)'\Sigma_Y^{-1}\boldsymbol{c}_Y(\boldsymbol{s}_0)\,, \tag{5}$$

which does not depend on $\boldsymbol{Y}$.

Scientific interest is in the process $Z(\cdot)$; hence, to predict $Z(\boldsymbol{s}_0)$ based on data $\boldsymbol{Z}$, classical prediction theory says the optimal predictor is obtained by minimizing the mean squared prediction error, $E(Z(\boldsymbol{s}_0)-p(\boldsymbol{Z};\boldsymbol{s}_0))^2$, with respect to predictor p. The theory further tells us that the best predictor is (e.g., Cressie, 1993, p. 108):

$$Z^*(\boldsymbol{s}_0) \;\equiv\; E(Z(\boldsymbol{s}_0)|\boldsymbol{Z})\,; \quad \boldsymbol{s}_0 \in D\,. \tag{6}$$

Calculation of (6) is not always possible, which explains why geostatisticians compromise with the best *linear* predictor. Because $Z(\cdot)$ is lognormal, it is unwise to use such a compromise here; in what follows, we evaluate (6). Since the conditional distribution of $Y(\boldsymbol{s}_0)|\boldsymbol{Y}$ is normal, then from (4) and (5),

$$\begin{aligned} Z^*(\boldsymbol{s}_0) \;&=\; \exp\{E(Y(\boldsymbol{s}_0)|\boldsymbol{Y}) + (1/2)\mathrm{var}(Y(\boldsymbol{s}_0)|\boldsymbol{Y})\}\,, \\ &=\; \exp\{Y^*(\boldsymbol{s}_0) + (1/2)C_Y(\boldsymbol{s}_0,\boldsymbol{s}_0) - (1/2)\boldsymbol{c}_Y(\boldsymbol{s}_0)'\Sigma_Y^{-1}\boldsymbol{c}_Y(\boldsymbol{s}_0)\} \\ &=\; \exp\{Y^*(\boldsymbol{s}_0) + (1/2)\mathrm{var}(Y(\boldsymbol{s}_0)) - (1/2)\mathrm{var}(Y^*(\boldsymbol{s}_0))\}\,. \end{aligned} \tag{7}$$

Clearly, the optimal predictor (7) is loglinear in the data and unbiased.

Recall that eventual interest is in the spatial prediction of block values $Z(B)$. The mean squared prediction error of a predictor $p(\boldsymbol{Z};B)$ of $Z(B)$ is $E(Z(B) - p(\boldsymbol{Z};B))^2$; $B \subset D$, and its minimization with respect to p yields the optimal predictor, $Z^*(B) = E(Z(B)|\boldsymbol{Z})$. Thus, $Z^*(B) = \int_B E(Z(\boldsymbol{u})|\boldsymbol{Z})d\boldsymbol{u}/|B| = \int_B Z^*(\boldsymbol{u})d\boldsymbol{u}/|B|$, where $Z^*(\cdot)$ is given by (6). That is, the optimal block predictor is,

$$Z^*(B) = \int_B \exp\{Y^*(\boldsymbol{u}) + (1/2)C_Y(\boldsymbol{u},\boldsymbol{u}) - (1/2)\boldsymbol{c}_Y(\boldsymbol{u})'\Sigma_Y^{-1}\boldsymbol{c}_Y(\boldsymbol{u})\}d\boldsymbol{u}/|B|\,. \tag{8}$$

The predictor $Z^*(B)$ was not used in the past because of a comment from Matheron (1974) that it "is too heavy to be used effectively in practice"; Rivoirard (1990) says it "would be possible but difficult" to compute, but Cressie (1993, p. 136) proposes it without comment about difficulties. In fact, vast increases in computing power in recent times has made quadrature of $\{Z^*(\boldsymbol{u}) \colon \boldsymbol{u} \in B\}$ given by (7), an easy computing exercise.

We now develop a second predictor of $Z(B)$. Suppose for the moment that we wish to predict $Y(B) \equiv \int_B Y(\boldsymbol{u})d\boldsymbol{u}/|B|$ based on data (on the log scale) $\boldsymbol{Y}$. The optimal spatial predictor is

$$Y^*(B) \;\equiv\; E(Y(B)|\boldsymbol{Y}) = \mu_Y(B) + \boldsymbol{c}_Y(B)'\Sigma_Y^{-1}(\boldsymbol{Y} - \boldsymbol{\mu}_Y)\,, \tag{9}$$

where $\mu_Y(B) \equiv \int_B \mu_Y(\boldsymbol{u})d\boldsymbol{u}/|B|$, $\boldsymbol{c}_Y(B) \equiv (C_Y(B,\boldsymbol{s}_1),\ldots,C_Y(B,\boldsymbol{s}_n))'$, and $C_Y(B,\boldsymbol{v}) \equiv \int_B C_Y(\boldsymbol{u},\boldsymbol{v})d\boldsymbol{u}/|B|$. One possible ad hoc block predictor is $Z^+(B) \equiv \exp\{Y^*(B)+k^+\}$, where k^+ is an adjustment for bias.

It is straightforward to see that $\exp\{Y^*(B)\}$ is lognormal, and hence

$$\begin{aligned} E(\exp\{Y^*(B)\}) &= \exp\{E(Y^*(B)) + (1/2)\mathrm{var}(Y^*(B))\} \\ &= \exp\{\int_B \mu_Y(\boldsymbol{u})d\boldsymbol{u}/|B| + (1/2)\boldsymbol{c}_Y(B)'\Sigma_Y^{-1}\boldsymbol{c}_Y(B)\}\,. \end{aligned}$$

Now, the mean of the predictand $Z(B)$ is,

$$\begin{aligned} E(Z(B)) &= \int_B E(\exp\{Y(\boldsymbol{u})\})d\boldsymbol{u}/|B| \\ &= \int_B \exp\{\mu_Y(\boldsymbol{u}) + (1/2)C_Y(\boldsymbol{u},\boldsymbol{u})\}d\boldsymbol{u}/|B|\,, \end{aligned}$$

from which the appropriate bias adjustment can be found. That is, an unbiased predictor of $Z(B)$ based on $Y^*(B)$ is:

$$\begin{aligned} Z^+(B) &= \exp\{Y^*(B) - \int_B \mu_Y(\boldsymbol{u})d\boldsymbol{u}/|B| - (1/2)\boldsymbol{c}_Y(B)'\Sigma_Y^{-1}\boldsymbol{c}_Y(B)\} \qquad (10)\\ &\quad \times \int \exp\{\mu_Y(\boldsymbol{u}) + (1/2)C_Y(\boldsymbol{u},\boldsymbol{u})\}d\boldsymbol{u}/|B|\,. \end{aligned}$$

We now give the block predictor $Z^@(B)$ based on a well known permanence approximation and show that it is very closely related to $Z^+(B)$. In its most general form, Cressie (2004) shows that

$$\begin{aligned} Z^@(B) &= \exp\{Y^*(B) + (1/2)\int_B C_Y(\boldsymbol{u},\boldsymbol{u})d\boldsymbol{u}/|B| \qquad (11)\\ &\quad +(1/2)\int_B (\mu_Y(\boldsymbol{u}) - \mu_Y(B))^2 d\boldsymbol{u}/|B| - (1/2)\boldsymbol{c}_Y(B)'\Sigma_Y^{-1}\boldsymbol{c}_Y(B)\}\,. \end{aligned}$$

Comparing the predictors (10) and (11), we see why (11) based on the permanence approximation is only approximately unbiased. In fact,

$$\begin{aligned} Z^@(B) &= Z^+(B)[\exp\{\int_B \mu_Y(\boldsymbol{u})d\boldsymbol{u}/|B| + (1/2)\int_B (\mu_Y(\boldsymbol{u}) - \mu_Y(B))^2 d\boldsymbol{u}/|B| \\ &\quad +(1/2)\int_B C_Y(\boldsymbol{u},\boldsymbol{u})d\boldsymbol{u}/|B|\}/\int_B \exp\{\mu_Y(\boldsymbol{u}) + (1/2)C_Y(\boldsymbol{u},\boldsymbol{u})\}d\boldsymbol{u}/|B|] \\ &\equiv Z^+(B)c_B\,. \end{aligned}$$

Now $Z^+(B)$ is unbiased for $Z(B)$; hence, when the factor c_B is 1, the block predictor $Z^@(B)$ based on the permanence approximation is also unbiased. As an example, suppose $\mu_Y(\boldsymbol{u}) \equiv \mu_Y$, a constant. Then the factor is:

$$c_B = \exp\{(1/2)\int_B C_Y(\boldsymbol{u},\boldsymbol{u})d\boldsymbol{u}/|B|\}/\int \exp\{(1/2)C_Y(\boldsymbol{u},\boldsymbol{u})\}d\boldsymbol{u}/|B|\,,$$

which is always ≤ 1, by Jensen's inequality. That is, for a constant mean function, $Z^{@}(B)$ has negative (or zero) bias. Suppose now that $\mu_Y(\boldsymbol{u}) \equiv \mu_Y$ *and* $C_Y(\boldsymbol{u}, \boldsymbol{u}) \equiv \sigma_Y^2$, which occurs whenever the Y process is second-order stationary. Then $c_B = 1$, the permanence-approximation-based predictor $Z^{@}(B)$ is equal to $Z^+(B)$, and hence $Z^{@}(B)$ is unbiased.

Although $Z^{@}(B)$ is the lognormal-kriging predictor that has been used traditionally (Cressie, 2004), it makes perfect sense to use the exactly unbiased predictor $Z^+(B)$ from now on. This can then be compared to the optimal predictor $Z^*(B)$ given by (8), to gauge how inefficient $Z^+(B)$ is. All of this assumes that the mean function is known; in the next subsection we consider an unknown mean.

2.2 UNKNOWN MEAN (COVARIANCE FUNCTION ASSUMED KNOWN)

In this section, and thereafter, we assume $\mu_Y(\cdot) \equiv \mu_Y$, a constant independent of location. Generalizaton to $\mu_Y(\cdot) = \boldsymbol{x}(\cdot)'\boldsymbol{\beta}$ can be achieved in a manner similar to that of Cressie (2004). We return to the problem of predicting $Z(\boldsymbol{s}_0)$ and note that $Z^*(\boldsymbol{s}_0)$ given by (7) depends on μ_Y through $Y^*(\boldsymbol{s}_0) = \mu_Y + \boldsymbol{c}_Y(\boldsymbol{s}_0)'\Sigma_Y^{-1}(\boldsymbol{Y} - \mu_Y\boldsymbol{1})$, where $\boldsymbol{1} \equiv (1, \ldots, 1)'$ is an $(n \times 1)$ vector of 1*s*. The simple-kriging predictor $Y^*(\boldsymbol{s}_0)$ becomes an ordinary-kriging predictor when the generalized least squares estimator for μ, $\widehat{\mu}_Y \equiv (\boldsymbol{1}'\Sigma_Y^{-1}\boldsymbol{1})^{-1}\boldsymbol{1}'\Sigma_Y^{-1}\boldsymbol{Y}$, is plugged in for the unknown μ_Y (Cressie, 1993, p. 173). That is, $\widehat{Y}(\boldsymbol{s}_0) = \widehat{\mu}_Y + \boldsymbol{c}_Y(\boldsymbol{s}_0)'\Sigma_Y^{-1}(\boldsymbol{Y} - \widehat{\mu}_Y\boldsymbol{1})$. Following the principal that the predictor should be unbiased on the original scale, we obtain the unbiased predictor (Matheron, 1974; Journel, 1980; Rivoirard, 1990; Cressie, 1993, p. 135),

$$\begin{aligned}\check{Z}(\boldsymbol{s}_0) &\equiv \exp\{\widehat{Y}(\boldsymbol{s}_0) + (1/2)\mathrm{var}(Y(\boldsymbol{s}_0)) - (1/2)\mathrm{var}(\widehat{Y}(\boldsymbol{s}_0))\} \\ &= \exp\{\widehat{Y}(\boldsymbol{s}_0) + (1/2)\sigma_{Y,k}^2(\boldsymbol{s}_0) - m(\boldsymbol{s}_0)\}\,, \qquad (12)\end{aligned}$$

where $\widehat{Y}(\boldsymbol{s}_0) \equiv \Sigma_{i=1}^n \lambda_i(\boldsymbol{s}_0)\boldsymbol{Y}(\boldsymbol{s}_0)$ is the ordinary-kriging predictor; $\boldsymbol{\lambda}(\boldsymbol{s}_0) \equiv (\lambda_1(\boldsymbol{s}_0), \ldots, \lambda_n(\boldsymbol{s}_0))'$ and $m(\boldsymbol{s}_0)$ solve the ordinary-kriging equations,

$$\Sigma_Y\boldsymbol{\lambda}(\boldsymbol{s}_0) = \boldsymbol{c}_Y(\boldsymbol{s}_0) + \boldsymbol{1}m(\boldsymbol{s}_0)\,, \qquad \boldsymbol{1}'\boldsymbol{\lambda}(\boldsymbol{s}_0) = 1\,;$$

and the kriging variance is $\sigma_{Y,k}^2(\boldsymbol{s}_0) = C_Y(\boldsymbol{s}_0, \boldsymbol{s}_0) - \boldsymbol{\lambda}(\boldsymbol{s}_0)'\boldsymbol{c}_Y(\boldsymbol{s}_0) + m(\boldsymbol{s}_0)$.

We define the optimal-prediction-based (o-p-b) predictor to be:

$$\check{Z}(B) \equiv \int_B \check{Z}(\boldsymbol{u})d\boldsymbol{u}/|B|\,, \qquad (13)$$

where $\check{Z}(\cdot)$ is given by (12). This will be compared with a (bias-adjusted) permanence-approximation-based (p-a-b) predictor, which we now derive.

The optimal block-kriging predictor of $Y(B)$ when μ_Y is known is given by (9). By substituting in the generalized least squares estimator, $\widehat{\mu}_Y = (\boldsymbol{1}'\Sigma_Y^{-1}\boldsymbol{1})^{-1}$ $(\boldsymbol{1}'\Sigma_Y^{-1}\boldsymbol{Y})$, we obtain the ordinary-block-kriging predictor, $\widehat{Y}(B) = \widehat{\mu}_Y + \boldsymbol{c}_Y(B)'\Sigma_Y^{-1}$ $(\boldsymbol{Y} - \widehat{\mu}_Y\boldsymbol{1}) \equiv \boldsymbol{\lambda}(B)'\boldsymbol{Y}$. Since $\widehat{Y}(B)$ is normal, then

$$\begin{aligned}E(\exp\{\widehat{Y}(B)\}) &= \exp\{E(\widehat{Y}(B)) + (1/2)\mathrm{var}(\widehat{Y}(B))\} \\ &= \exp\{\mu_Y + (1/2)\boldsymbol{\lambda}(B)'\Sigma_Y\boldsymbol{\lambda}(B)\}\,,\end{aligned}$$

and hence the analogous expression to (10) yields the p-a-b predictor:

$$\begin{aligned}
\widehat{Z}(B) &= \exp\{\widehat{Y}(B) - (1/2)\boldsymbol{\lambda}(B)'\Sigma_Y\boldsymbol{\lambda}(B)\} \times \int \exp\{(1/2)C_Y(\boldsymbol{u},\boldsymbol{u})\}d\boldsymbol{u}/|B| \\
&= \exp\{\widehat{Y}(B) - (1/2)\int_B\int_B C_Y(\boldsymbol{u},\boldsymbol{v})d\boldsymbol{u}d\boldsymbol{v}/|B|^2 + (1/2)\sigma^2_{Y,k}(B) - m(B)\} \\
&\qquad\qquad \times \int \exp\{(1/2)C_Y(\boldsymbol{u},\boldsymbol{u})\}d\boldsymbol{u}/|B| \qquad (14) \\
&\equiv \exp\{\widehat{Y}(B) + \widehat{k}\}\,,
\end{aligned}$$

where $\boldsymbol{\lambda}(B)$ and $m(B)$ solve the ordinary-block-kriging equations,

$$\Sigma_Y\boldsymbol{\lambda}(B) = \boldsymbol{c}_Y(B) + \mathbf{1}m(B)\,, \qquad \mathbf{1}'\boldsymbol{\lambda}(B) = 1\,;$$

and the kriging variance is $\sigma^2_{Y,k}(B) = \int_B\int_B C(\boldsymbol{u},\boldsymbol{v})d\boldsymbol{u}d\boldsymbol{v}/|B|^2 - \boldsymbol{\lambda}(B)'\boldsymbol{c}_Y(B) + m(B)$. A proof of the equality that results in (14) is given in Cressie (2004).

In the next section, we compare the o-p-b predictor (13) with the p-a-b predictor (14). Both are unbiased and hence the comparison is through mean squared prediction errors.

3 Comparison of predictors

The comparison of $\check{Z}(B)$ given by (13) and $\widehat{Z}(B)$ given by (14), via mean squared prediction errors, has a theoretical component and a simulation component.

3.1 THEORETICAL EXPRESSIONS

Cressie (2004) derived the mean squared prediction error of $\check{Z}(B)$ through:

$$E(Z(B) - \check{Z}(B))^2 = \int_B\int_B \mathrm{cov}(Z(\boldsymbol{u}) - \check{Z}(\boldsymbol{u}), Z(\boldsymbol{v}) - \check{Z}(\boldsymbol{v}))d\boldsymbol{u}d\boldsymbol{v}/|B|^2\,, \qquad (15)$$

where the integrand of (15) is given by

$$(\exp\{\mu_Y + (1/2)C_Y(\boldsymbol{u},\boldsymbol{u})\})(\exp\{\mu_Y + (1/2)C_Y(\boldsymbol{v},\boldsymbol{v})\})(a - b - c + d)\,,$$

and

$$\begin{aligned}
a &= \exp\{C_Y(\boldsymbol{u},\boldsymbol{v})\}\,, \\
b &= \exp\{(\boldsymbol{c}_Y(\boldsymbol{u}) + \mathbf{1}m(\boldsymbol{u}))'\Sigma_Y^{-1}\boldsymbol{c}_Y(\boldsymbol{v})\}\,, \\
c &= \exp\{(\boldsymbol{c}_Y(\boldsymbol{v}) + \mathbf{1}m(\boldsymbol{v}))'\Sigma_Y^{-1}\boldsymbol{c}_Y(\boldsymbol{u})\}\,, \\
d &= \exp\{(\boldsymbol{c}_Y(\boldsymbol{u}) + \mathbf{1}m(\boldsymbol{u}))'\Sigma_Y^{-1}(\boldsymbol{c}_Y(\boldsymbol{v}) + \mathbf{1}m(\boldsymbol{v}))\}\,.
\end{aligned}$$

In practice, the double integral in (15) is approximated with a double summation.

Using a similar derivation to the one found in Cressie (2004), the mean squared prediction error of (14) is seen to be

$$\begin{aligned} E(Z(B)-\widehat{Z}(B))^2 &= \int_B\int_B \operatorname{cov}(\exp\{Y(\boldsymbol{u})\},\exp\{Y(\boldsymbol{v})\})d\boldsymbol{u}d\boldsymbol{v}/|B|^2 \\ &\qquad + \operatorname{var}(\widehat{Z}(B)) - 2\int_B \operatorname{cov}(\exp\{Y(\boldsymbol{u})\},\widehat{Z}(B))d\boldsymbol{u}/|B| \\ &\equiv \Big(\int_B\int_B f\, d\boldsymbol{u}d\boldsymbol{v}/|B|^2\Big) + g - 2\Big(\int_B h\, d\boldsymbol{u}/|B|\Big)\,, \end{aligned} \tag{16}$$

where

$$\begin{aligned} f &= (\exp\{\mu_Y + (1/2)C_Y(\boldsymbol{u},\boldsymbol{u})\})(\exp\{\mu_Y + (1/2)C_Y(\boldsymbol{v},\boldsymbol{v})\})(\exp\{C_Y(\boldsymbol{u},\boldsymbol{v})\}-1) \\ g &= (\exp\{2\widehat{k}\})(\exp\{2\mu_Y + \boldsymbol{\lambda}(B)'\Sigma_Y\boldsymbol{\lambda}(B)\})(\exp\{\boldsymbol{\lambda}(B)'\Sigma_Y\boldsymbol{\lambda}(B)\}-1) \\ h &= (\exp\{\widehat{k}\})(\exp\{\mu_Y + (1/2)C_Y(\boldsymbol{u},\boldsymbol{u})\})(\exp\{\mu_Y + (1/2)\boldsymbol{\lambda}(B)'\Sigma_Y\boldsymbol{\lambda}(B)\}) \\ &\qquad \cdot (\exp\{\boldsymbol{\lambda}(B)'\boldsymbol{c}_Y(\boldsymbol{u})\}-1)\,, \end{aligned}$$

and $\exp(\widehat{k}) = \exp\{-(1/2)\boldsymbol{\lambda}(B)'\Sigma_Y\boldsymbol{\lambda}(B)\} \times \int_B \exp\{(1/2)C_Y(\boldsymbol{u},\boldsymbol{u})\}d\boldsymbol{u}/|B|$.

3.2 EMPIRICAL COMPARISON VIA SIMULATION

A simulation experiment was conducted in order to compare the two lognormal kriging predictors, $\check{Z}(B)$ given by (13) and $\widehat{Z}(B)$ given by (14). We expect (13) to have better performance than (14), since it is developed from the optimal predictor. Still, the question remains as to how close the two predictors are in practice, since the permanence approximation has been used a lot in past applications.

We generated a spatially dependent Gaussian process $Y(\cdot)$ on a 32×32 square with 33×33 nodes, each one unit apart; see Figure 1. The Gaussian process had $\mu_Y = 0$, and an isotropic covariance function $C_Y(\boldsymbol{u},\boldsymbol{v}) = C_Y^{(0)}(\|\boldsymbol{u}-\boldsymbol{v}\|) \geq 0$, given by the spatial moving average described in Cressie and Pavlicová (2002). Different parameters were varied.

- Sill: $C_Y^{(0)}(0) \equiv \sigma_Y^2 \in \{0.1, 0.7, 2.5\}$.
- Nugget effect: $\lim_{h\to 0}\{C_Y^{(0)}(0) - C_Y^{(0)}(h)\}/C_Y^{(0)}(0) \equiv \nu \in \{0\%, 10\%, 30\%, 50\%\}$.
- Range: $R \equiv \arg\inf\{h\colon C_Y(h') = 0, h' \geq h\} \in \{8, 16, 32, 64\}$.

The sill values σ_Y^2 were chosen to give a representative range of coefficients of variation, $CV \equiv (\exp\{\sigma_Y^2\} - 1)^{1/2} \in \{.32, 1.01, 3.34\}$. The nugget effect ν is anywhere up to 50% of the sill, and the linear dimension of the 32×32 square is anywhere between four times (weak spatial dependence) and half (very strong spatial dependence) the range R. Although observations on $Y(\cdot)$ were simulated at each grid node $\{\boldsymbol{u}_i\colon i = 1,\ldots,33\times 33\}$, only a subset $\{\boldsymbol{s}_i\colon i = 1,\ldots,n\}$ were used to generate data for the experiment.

- Data: $Z(\boldsymbol{s}_i) \equiv \exp\{Y(\boldsymbol{s}_i)\}$; $i = 1,\ldots,n$, where $n \in \{4, 25, 81\}$ and the data were nested according to Figure 1. For $n = 4$, the data are 16 units apart; for $n = 25$, the data are 8 units apart; for $n = 81$, the data are 4 units apart.

Lognormal kriging is carried out on blocks of varying supports since the permanence approximation is likely to be better for smaller blocks.

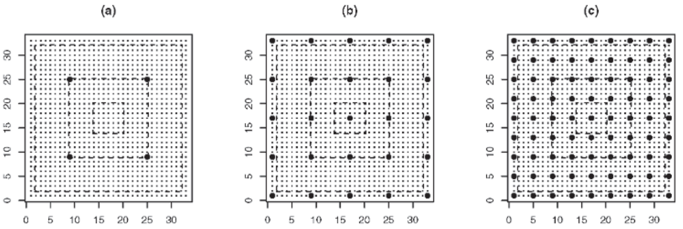


Figure 1. Grid upon which data are simulated; locations of data are shown: (a) 4 observed locations; (b) 25 observed locations; (c) 81 observed locations. The broken lines show 3 different block sizes (small, medium, large).

- Support: Predict $Z(B)$ on blocks $B \in \{2\times 2, 4\times 4, 6\times 6, \ldots, 32\times 32\}$, centered on the center of the 33×33 grid $\{\boldsymbol{u}_i\}$ and nested. Small support (6×6), medium support (16×16), and large support (30×30) are featured; see Figure 1.

Finally, the responses of the experiment are based on the two fundamental properties of a predictor: bias and mean squared prediction error. For both predictors, $\check{Z}(B)$ given by (13) and $\widehat{Z}(B)$ given by (14), theory tells us that the bias should be zero. However, the predictors should differ in mean squared prediction error, where that for $\check{Z}(B)$ is expected to be smaller.

Let $Z^{(\ell)}(\cdot)$ denote the ℓ-th simulation of the log Gaussian process with specified σ_Y^2, ν, and R; $\ell = 1, \ldots, L$, and assume $Z^{\dagger}(B)$ is a generic predictor of $Z(B)$. Then define the mean squared prediction error of $Z^{\dagger}(B)$ as:

$$\text{MPSE} \;\equiv\; (1/L)\sum_{\ell=1}^{L}\{Z^{\dagger}(B) - Z(B)\}^2 \,, \tag{17}$$

where any integrals in $Z^{\dagger}(B)$ or $Z(B)$ are approximated as sums based on the finest grid spacing. The value $L = 6400$ was chosen to guarantee accuracy of results to the second decimal place, where that digit is conservatively plus or minus 2; see Aldworth and Cressie (1999) for the relevant calculations that determine L.

From the simulation experiment we conclude that:

- The average observed biases are extremely close to 0.
- The theoretical mean squared prediction errors and their empirical counterparts (MSPEs defined in (17)) are approximately equal; see Figure 2.
- The MSPE for $\check{Z}(B)$ is smaller than (and occasionally equal to, up to sampling error) the MSPE for $\widehat{Z}(B)$. That is, the spatial predictor $\check{Z}(B)$ is dominant over $\widehat{Z}(B)$; see Figure 3(a).

It is this latter result that we would like to explore at greater depth, since the improvement in efficiency obtained by using $\check{Z}(B)$ is not uniform over all combinations of the factors of the simulation experiment. The efficiency of the p-a-b predictor $\widehat{Z}(B)$ relative to the o-p-b predictor $\check{Z}(B)$ is defined as $E \equiv \check{MSPE}/\widehat{MSPE}$, where $\check{MSPE}$ ($\widehat{MSPE}$) is given by (17) with $Z^{\dagger} \equiv \check{Z}$ ($Z^{\dagger} \equiv \widehat{Z}$).

Figure 3(a) shows E for all combinations of factor levels and the dominance of $\check{Z}(B)$ over $\widehat{Z}(B)$ is striking. Figure 3(b) shows that the efficiency of $\widehat{Z}(B)$ decreases

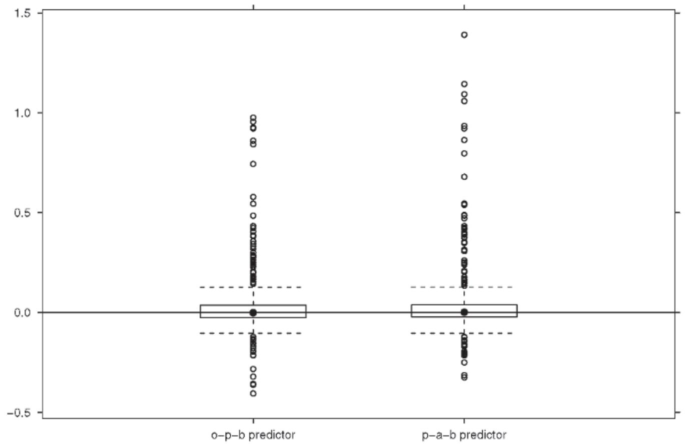


Figure 2. Comparison of theoretical mean squared prediction errors and MSPEs given by (17), for the optimal-prediction-based (o-p-b) predictor $\widehat{Z}(B)$ and the permanence-approximation-based (p-a-b) predictor $\widehat{Z}(B)$. Shown is the ratio of theoretical over MSPE, minus 1, displayed over all combinations of factor levels.

as the sill σ_Y^2 increases; that is, the more skewed the lognormal distribution, the greater are the potential gains in efficiency using $\check{Z}(B)$. Figure 3(c) shows that the efficiency of $\widehat{Z}(B)$ increases as the nugget effect ν increases; that is, as the spatial dependence gets weaker, $\check{Z}(B)$ is not as dominant over $\widehat{Z}(B)$. A plot of E broken down by range (not shown here) reinforces this observation; for small R (weaker spatial dependence), $\check{Z}(B)$ is not as dominant over $\widehat{Z}(B)$. From Figure 3(d), we see that $\check{Z}(B)$ dominates over $\widehat{Z}(B)$ when data are closer together (in units of range) and is less dominant when they are far apart. That is, the more nearby the spatial data are, the better $\check{Z}(B)$ is able to use that information.

4 Conclusion

We recommend unequivocably that for lognormal kriging, the optimal-prediction-based predictor $\check{Z}(B)$ given by (13) be used. It is unbiased with mean squared prediction error (i.e., kriging variance) given by (15).

References

Aitchison, J. and Brown, J. A. C. *The Lognormal Distribution*, Cambridge University Press, 1957.

Aldworth, J. and Cressie, N. Sampling designs and prediction methods for Gaussian spatial processes, in *Multivariate Design and Sampling*, Ghosh, S. (ed.), Marcel Dekker, NY, p. 1-54.

Cressie, N. *Statistics for Spatial Data (Revised Edition)*, Wiley, NY, 1993.

Cressie, N. Block kriging for lognormal spatial processes. Department of Statistics Preprint No. 739, The Ohio State University, 2004.

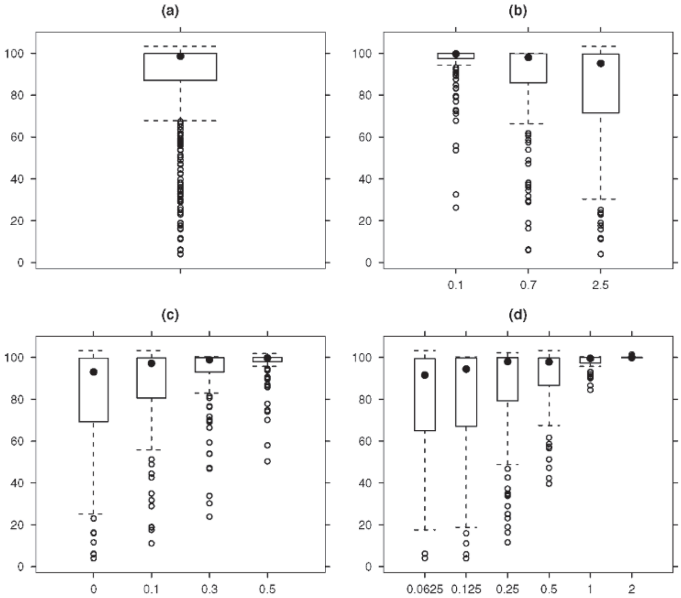


Figure 3. Efficiency E of $\widehat{Z}(B)$ relative to $\check{Z}(B)$: (a) for all factors combined; (b) broken down by sill; (c) broken down by nugget effect; (d) broken down by the distance between locations of observed data (in units of range).

Cressie, N. and Pavlicová, M., 2002. Calibrated spatial moving average simulations, *Statistical Modelling*, vol. 2, p. 267-279.

Dowd, P. A. Lognormal kriging - The general case, *Mathematical Geology*, vol. 14, 1982, p. 475-499.

Journel, A. G. The lognormal approach to predicting local distributions of selective mining unit grades, *Mathematical Geology*, vol. 12, 1980, p. 285-303.

Marechal, A. Krigeage normal et lognormal, Centre de Morphologie Mathematique, Ecole des Mines de Paris, Publication N-376.

Matheron, G. Traite de Geostatistique Appliquee, Tome 1, *Memoires du Bureau de Recherche Geologiques et Minieres*, no. 14, ed. Technip, Paris, 1962.

Matheron, G. Effet proportionnel et lognormalite ou: Le retour du serpent de mer, Centre de Morphologie Mathematique, Ecole des Mines de Paris, Publication N-374.

Rendu, J.-M. Normal and lognormal estimation, *Mathematical Geology*, vol. 11, 1979, p. 407-422.

Rivoirard, J. A review of lognormal estimators for in situ reserves, *Mathematical Geology*, vol. 22, 1990, p. 213-221.

Roth, C. Is lognormal kriging suitable for local estimation?, *Mathematical Geology*, vol. 30, 1988, p. 999-1009.

EVALUATING INFORMATION REDUNDANCY THROUGH THE TAU MODEL

SUNDERRAJAN KRISHNAN, ALEXANDRE BOUCHER and ANDRE G. JOURNEL
Department of Geological & Environmental Sciences, Stanford University, California

Abstract.
The general problem of data integration is expressed into that of combining individual probabilistic information into a joint posterior probability. Any such combination of information necessarily requires taking into account redundancy within the information utilized. It is shown that the tau model (Journel 2002) can provide an exact analytical representation of such combination. The tau weights express data redundancy for each specific sequence of data conditioning.

Instead of using this exact definition of the tau weights, a more practical calibration-based method is proposed. The method requires a prior ranking of the data based on their information content, then the tau weights are approximated by a function of the correlation of each datum with the single most informative one. Such calibration would require training information in the form of joint vectorial data.

The tau model can also be expressed as a log-linear estimator of the distance to the unknown event. This definition requires considering the distances (or equivalently the odds ratios) as random variables themselves. An application to binary data is presented.

Keywords: Combining information, tau model, data redundancy, posterior probability, conditional independence

1 Statement of problem

Combining information from different sources is a difficult problem occurring over many different disciplines. Consider that we wish to assess our knowledge about an event A. Here A could be as complex as presence/absence of a set of connected fractures close to a well or it could be the binary event that the average porosity of a given region in the subsurface is lesser than a given threshold value. Typically, we get information about the unknown event from, say, n different sources namely D_1, D_2, ..., D_n. Each datum event D_i can be quite complex involving different variables and multiple sample locations in space.

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 1037–1046.

The conditional probability $P(A|D_i)$ is a convenient way of expressing the information conveyed by the single datum event D_i. The extreme values $P(A|D_i)$ $= 0$ or 1 correspond to decisive information about A as provided by D_i. Denote the prior probability of A occurring as $P(A)$, i.e., as obtained from some prior information available "prior" to getting any of the data D_i.

The problem of combining information can then be stated as:

Given the prior information $P(A)$ and the n data D_i, $i = 1, \dots n$, how can we combine the single-datum conditional probabilities $P(A|D_i)$ into a posterior probability $P(A|D_1, \dots, D_n)$, conditioned to all data events taken together.

An excellent review of techniques addressing aspects of this problem is given by Genest and Zidek(1986). Many of these techniques call for some sort of independence assumption which skips the problem of data redundancy. Here, we develop the tau model to explicitly account for such redundancy, presenting an interpretation of the tau weights and proposing methods for calibrating these weights.

2 The tau model

Define the following data probability ratios $x_0, x_1, \dots, x_n$ and the target ratio x, all valued in $[0, \infty]$, as:

$$x_0 = \tfrac{1-P(A)}{P(A)} \ , x_1 = \tfrac{1-P(A|D_1)}{P(A|D_1)}, \ \dots, x_n = \tfrac{1-P(A|D_n)}{P(A|D_n)} \ \text{ and } \ x = \tfrac{1-P(A|D_1,\dots,D_n)}{P(A|D_1,\dots,D_n)},$$

Note that $P(A|D_i) = 1$ gives $x_i = 0$ and $P(A|D_i) = 0$ gives $x_i = \infty$, therefore these ratios can be interpreted as distances to the unknown A occurring. They are also equal to the inverse odds ratio. It is then required to compute the distance x of the joint data event to the unknown from knowledge of the individual distances $x_0, x_1, \dots, x_n$. The tau model is then stated as (Bordley 1982; Journel 2002):

$$\frac{x}{x_0} = \prod_{i=1}^{n} \Big(\frac{x_i}{x_0}\Big)^{\tau_i} \tag{1}$$

with:

$$P(A|D_1, \dots, D_n) = \tfrac{1}{1+x} \ \epsilon \ [0, 1]$$

The tau weights $\tau_i \epsilon [-\infty, \infty]$, $i = 1, \dots, n$ account for the redundancy between the n data. One property of this model is that if $P(A|D_i) = 0$ or 1, then $P(A|D_1, \dots, D_n) = 0$ or 1. Individual certainty implies overall certainty.

2.1 TAU WEIGHTS AND SEQUENTIAL DATA REDUNDANCY

Consider a specific sequence s of the n data: $D_1^{(s)}, \dots, D_n^{(s)}$. With n data there are a total number of $S = n!$ such sequences. From the definition of conditional probability and its decomposition, one can write:

$$P(A|D_1^{(s)}, \dots, D_n^{(s)}) = \frac{P(D_1^{(s)}) \ P(A \ D_1^{(s)}) \ P(D_2^{(s)} \ A,D_1^{(s)}) \dots P(D_n^{(s)} \ A,D_1^{(s)},\dots,D_{n-1}^{(s)})}{P(D_1^{(s)},\dots,D_n^{(s)})}$$

and similarly for the conditional probability $P(\tilde{A}|D_1^{(s)},\ldots,D_n^{(s)})$ of the complementary event
$\tilde{A} = \text{non } A$.

We then get the ratio:

$$\frac{P(\tilde{A}\ D_1^{(s)},...,D_n^{(s)})}{P(A\ D_1^{(s)},...,D_n^{(s)})} = \frac{P(\tilde{A}\ D_1^{(s)})}{P(A\ D_1^{(s)})}\frac{P(D_2^{(s)}\ \tilde{A},D_1^{(s)})}{P(D_2^{(s)}\ A,D_1^{(s)})}\cdots\frac{P(D_n^{(s)}\ \tilde{A},D_1^{(s)},...,D_{n-1}^{(s)})}{P(D_n^{(s)}\ A,D_1^{(s)},...,D_{n-1}^{(s)})}$$

Using the previous definitions of the distances x, x_1, we get:

$$x = x_1^{(s)}\frac{P(D_2^{(s)}|\tilde{A},D_1^{(s)})}{P(D_2^{(s)}|A,D_1^{(s)})}\cdots\frac{P(D_n^{(s)}|\tilde{A},D_1^{(s)},\ldots,D_{n-1}^{(s)})}{P(D_n^{(s)}|A,D_1^{(s)},\ldots,D_{n-1}^{(s)})} \tag{2}$$

The analytically exact expression (2) identifies the tau model (1) if each of the n individual likelihood ratios $\frac{P(D_i^{(s)}|\tilde{A},D_1^{(s)},...,D_{i-1}^{(s)})}{P(D_i^{(s)}|A,D_1^{(s)},...,D_{i-1}^{(s)})}$ are written in terms of the tau parameters as:

$$\frac{P(D_i^{(s)}|\tilde{A},D_1^{(s)},\ldots,D_{i-1}^{(s)})}{P(D_i^{(s)}|A,D_1^{(s)},\ldots,D_{i-1}^{(s)})} = \Big[\frac{P(D_i^{(s)}|\tilde{A}}{P(D_i^{(s)}|A)}\Big]^{\tau_i^{(s)}},\quad i=2,\ldots,n \tag{3}$$

Thus:

$$\tau_1^{(s)} = 1,\quad \tau_i^{(s)} = \frac{log(\frac{P(D_i^{(s)}|\tilde{A},D_1^{(s)},...,D_{i-1}^{(s)})}{P(D_i^{(s)}|A,D_1^{(s)},...,D_{i-1}^{(s)})})}{log(\frac{P(D_i^{(s)}|\tilde{A})}{P(D_i^{(s)}|A)})}\ \epsilon\ [-\infty,+\infty],\quad i=2,\ldots,n \tag{4}$$

The tau parameters are expressed as ratios of log of likelihood ratios, and they depend on the specific sequence of data conditioning s. Consider first the second tau parameter $\tau_2^{(s)}$: the denominator in expression (4) measures what is the sensitivity of $D_2^{(s)}$ to A changing from any value to its complement. Note that this ratio is specific to a particular realization of the variables $A = a$ and $D_2^{(s)} = d_2$. That denominator is assumed non-zero, that is $D_2^{(s)}$ distinguishes A from $\tilde{A}$, otherwise the information $D_2^{(s)}$ would simply be ignored. The numerator in (4) can be seen as the same sensitivity of $D_2^{(s)}$ to A but now in presence of $D_1^{(s)}$. Therefore, the tau weight $\tau_2^{(s)}$ can be seen as the change in sensitivity of $D_2^{(s)}$ to A brought by knowledge of the first datum $D_1^{(s)}$. This weight is specific to the ordering which sets $D_1^{(s)}$ as the first conditioning datum. Changing the ordering will give a different weight.

More generally, $\tau_i^{(s)}$ measures the change in sensitivity of datum $D_i^{(s)}$ to A brought by knowledge of all previous data $D_1^{(s)},\ldots,D_{i-1}^{(s)}$ in the sequence s. Substituting expressions (3) into (2) for $i = 2,\ldots,n$, we get the sequence-specific tau model:

$$x = x_1^{(s)} \left(\frac{x_2^{(s)}}{x_0}\right)^{\tau_2^{(s)}} \cdots \left(\frac{x_n^{(s)}}{x_0}\right)^{\tau_n^{(s)}} \tag{5}$$

Expression (5) identifies the exact expression for the conditional probability $P(A|D_1^{(s)},\ldots,D_n^{(s)})$. The tau weights are dependent on the specific outcome value of the random variables $A, D_1^{(s)},\ldots,D_n^{(s)}$ and are also dependent on the specific sequence s of data conditioning.

2.2 INTERPRETING THE TAU WEIGHTS

Expression (4) gives the ith tau weight $\tau_i^{(s)}$ dependent on the previous $(i-1)$ data $D_1^{(s)},\ldots,D_{i-1}^{(s)}$.

Consider the following values for that weight:

- $\tau_i^{(s)} = 1$:

$\tau_i^{(s)}$ equal to 1 requires that the ratios $\frac{P(D_i^{(s)}|\tilde{A},D_1^{(s)},\ldots,D_{i-1}^{(s)})}{P(D_i^{(s)}|A,D_1^{(s)},\ldots,D_{i-1}^{(s)})}$ and $\frac{P(D_i^{(s)}|\tilde{A})}{P(D_i^{(s)}|A)}$ be equal to each other. One possibility is:

$P(D_i^{(s)}|\tilde{A}, D_1^{(s)},\ldots,D_{i-1}^{(s)}) = P(D_i^{(s)}|\tilde{A})$ and

$P(D_i^{(s)}|A, D_1^{(s)},\ldots,D_{i-1}^{(s)}) = P(D_i^{(s)}|A)$.

This means that datum $D_i^{(s)}$ is independent of the previous data in sequence s given A, i.e. conditional independence along the sequence s. However, in general, equality of these ratios does not require conditional independence, but only that: $\frac{P(D_i^{(s)}|\tilde{A},D_1^{(s)},\ldots,D_{i-1}^{(s)})}{P(D_i^{(s)}|\tilde{A})} = \frac{P(D_i^{(s)}|A,D_1^{(s)},\ldots,D_{i-1}^{(s)})}{P(D_i^{(s)}|A)} =$ some constant r. The value $r =$ 1 arises from conditional independence. Any other $r \,\epsilon\, [0,\infty]$ would also give $\tau_i^{(s)} =$ 1

- $\tau_i^{(s)} = 0$: incremental non-information

This means that the numerator in expression (4) is equal to 0 calling for:

$P(D_i^{(s)}|\tilde{A}, D_1^{(s)},\ldots,D_{i-1}^{(s)}) = P(D_i^{(s)}|A, D_1^{(s)},\ldots,D_{i-1}^{(s)})$,

i.e., $D_i^{(s)}$ is independent of A given the set of $(i-1)$ previous data. $D_i^{(s)}$ does not add anything towards knowledge of A. Note again that this does not mean that datum $D_i^{(s)}$ is independent of A. Another sequence s' could result in a non-zero weight.

- $\tau_i^{(s)} > 1,\ \tau_i^{(s)} \epsilon\ (0,1)$: amplifying/diminishing sensitivity

A weight greater than 1 implies that the sensitivity of datum $D_i^{(s)}$ to change in A is amplified by knowledge of the $(i-1)$ previous data. The reverse holds true for $\tau_i^{(s)} < 1$.

- $\tau_i^{(s)} < 0$: reversal of sensitivity

A negative sign for weight $\tau_i^{(s)}$ means that the numerator and denominator of expression (4) have opposite signs. This means that knowledge of the previous data reverses the sensitivity of datum $D_i^{(s)}$ to A when taken individually. The joint impact of $D_i^{(s)}$ is opposite to its individual impact. Again, this is true only for the particular sequence of previous data $D_1^{(s)}, \dots, D_{i-1}^{(s)}$.

As shown here, the tau weights arrive from a specific ordering of the data conditioning. All previous concepts - conditional independence, incremental non-information, amplifying/diminishing sensitivity and reversal of sensitivity - are function of the values taken by the previous data in the sequence. Therefore, they do not truly reflect the overall dependency between the variables. One could argue, however, that as long as a certain set of tau weights can reproduce the posterior probability closely enough, that is all that is required for practice.

However, in order to study the general impact of data redundancy on the tau weights, it is necessary to consider the different sets of tau weights resulting from different ordering of data. One possibility is to average the tau weights $\tau_i^{(s)}$ for each datum D_i over all possible sequences $s = 1, \dots, S$ to obtain a sequence-independent weight $\overline{\tau_i}$.

3 Computing the tau weights

Here, we address the practical question of computing the sequence-dependent tau weights $\tau_i^{(s)}$ using a calibration-based technique. The first step thus calls for determining a specific sequence of data. Note that any sequence is suitable as long as expression (4) can be evaluated.

One possibility is to base this ordering on the information provided by the data towards knowledge of A. The greater is that individual information content, the higher should be the rank of the datum. This requires a measure of information content for the data. Literature in information theory is abundant providing numerous such measures, for example the mutual information measure (McEliece, 2002). Here, we use a simple measure based on the deviation from the prior information $P(A)$.

The standardized information content $\zeta(D_i, A)$ proposed by Liu (2002) is:

$$\begin{array}{c}\zeta(D_i, A) = \\ \epsilon[0,1]\end{array} \begin{cases} \frac{P(A)-P(A|D_i)}{P(A)}, & \text{if } P(A|D_i) \leq P(A) \\ \\ \frac{P(A|D_i)-P(A)}{1-P(A)}, & \text{if } P(A|D_i) \geq P(A) \end{cases} \tag{6}$$

It assumes a linear variation of the information content, ranging from 0 when $P(A|D_i)$ is equal to the prior probability $P(A)$ to a maximum 1 at both extremes when $P(A|D_i)$ is equal to 0 or 1. Using this information measure, the n data can be ranked in decreasing order, so that D_1 is the most informative and D_n, the least informative datum.

Following from expression (5), the weight τ_1 is set to 1. As for τ_2, instead of the exact expression (4), we consider a heuristic approximation using the conditional correlation $\rho^2_{D_1,D_2|A}$ of D_2 and D_1 given A. Intuitively, the greater this correlation, the lesser should be the weight τ_2 given to the less informative data:

$$\tau_2 \sim 1 - \rho^2_{D_1,D_2|A}$$

Similarly, the third weight τ_3 would be such that:

$$\tau_3 \sim 1 - \rho^2_{D_3,(D_1,D_2)|A}$$

where $\rho^2_{D_3,(D_1,D_2)|A}$ is the conditional correlation of D_3 with the set (D_1, D_2) given A.

Continuing to the nth weight τ_n:

$$\tau_n \sim 1 - \rho^2_{D_n,(D_1,\ldots,D_{n-1})|A}$$

The conditional correlation $\rho^2_{D_n,(D_1,\ldots,D_{n-1})|A}$ would capture the redundancy of D_n with all previous data in the sequence. In practice, such conditional correlation would be difficult to obtain. Hence we propose to approximate it by the conditional correlation with the single most informative datum D_1, giving,

$$\tau_n \sim 1 - \rho^2_{D_n,D_1|A}$$

Providing a calibration parameter t, we can then write the weight τ_n as:

$$\tau_n = 1 - (\rho^2_{D_n,D_1|A})^{f(t)}$$

where $t \,\epsilon\, [0, 1]$ is a calibration parameter and $f(t) = Ln(1/(1 - t))$ is a scaling function. We use the Ln transform in order to map $t \,\epsilon [0, 1]$ into $f(t) \,\epsilon\, [0, \infty]$. This allows for a standardized parameter $t \,\epsilon [0, 1]$. Here, the correlations ρ^2 capture the redundancy between data as to informing A, and the parameter t rescales this relationship to fit to any available training data.

Summarizing this proposed calibration-based method, we first order the data in terms of their information content. Then, the most informative datum D_1 is given the maximum weight of $\tau_1 = 1$.

Any other datum D_i is given a weight τ_i given by:

$$\tau_i = 1 - (\rho^2_{D_i,D_1|A})^{f(t)} \;\epsilon\; [0, 1) \tag{7}$$

Note that these weights τ_i lie in $[0, 1)$ as opposed to the tau weights given in expression (4) which have no such restriction. Therefore, this calibration method should not be considered for cases where one ought to consider weights $\tau_i > 1$ (information amplification) and $\tau_i < 0$ (information reversal).

Expression (7) requires computation of all the $n - 1$ conditional correlations $\rho^2_{D_i,D_1|A}$ and calibration of the single parameter t. Such computation will require availability of some approximation of the $(n+1)$ variables $A, D_1, \ldots, D_n$ possibly under the form of joint (vectorial) training data.

4 Probability distances as random variables

The analysis in the previous sections can be seen as a Bayesian approach to combining probabilistic information. The prior information is sequentially updated by new data. At every stage we consider the worth of the new information in terms of its redundancy with all previous data. That is just one interpretation of the tau model. Going back to the definition of the probabilistic distances $x_0,\ x_i,\ x$, it is possible to consider these distances as random variables themselves. Rewriting the tau model (1) by taking logarithm, the log-posterior distance x is given by:

$$log\ x^* - log\ x_0 = \sum_{i=1}^{n} \tau_i [\ log\ x_i - log\ x_0]$$

First, we recognize that the conditional probability $P(A|D_i)$ can be derived from the joint distribution of the variables A and D_i. One can also view this conditional probability $P(A|D_i)$ as a function valued in $[0,1]$ of the RV D_i. Here, the discussion is limited to binary indicator variables $A, D_1, \ldots, D_n$ taking values 0/1, but the approach is general. This makes $P(A|D_i)$ a binary random function of D_i (and A) taking on two values $P(A|D_i = 1)$ with probability $P(D_i = 1)$ and $P(A|D_i = 0)$ with the complement probability $P(D_i = 0)$.

The same is true for any transform of the conditional probability $P(A|D_i)$, namely the distances x_i or the log-distances $y_i = log(x_i)$. Define the RV X_i as a binary variable taking on two values:

$x_i^+ = \frac{1-P(A|D_i=1)}{P(A|D_i=1)}$ with probability $P(D_i = 1)$ and
$x_i^- = \frac{1-P(A|D_i=0)}{P(A|D_i=0)}$ with probability $P(D_i = 0)$.

The RV Y_i can now be defined as $Y_i = log(X_i)$ taking on two values $y_i^+ = log(x_i^+)$ and $y_i^- = log(x_i^-)$.

Similarly, the target RV Y can be considered a function of D, taking on two values:

$Y = y^+ = \frac{1-P(A|D=1)}{P(A|D=1)}$ with probability $P(D = 1)$ and
$Y = y^- = \frac{1-P(A|D=0)}{P(A|D=0)}$ with the complement probability $P(D = 0)$. Here $P(D)$ is the joint probability of all n data occurring together, for example $D = \prod_{i=1}^{n} D_i = 1$, therefore D is a multiple-datum statistic.

Tau model as a kriging estimator

The previous interpretation of the log-distances allows to rewrite the tau model as an estimator of the target log-probability distance Y.

Taking logarithm on the tau model expression (1), we get:

$$Y^* - log(x_0) = \sum_{i=1}^{n} \tau_i (Y_i - log(x_0)) \tag{8}$$

Note that this is not a Simple Kriging (SK) expression since $E\{Y\} \neq log(x_0)$. However, a system similar to kriging can be developed for this estimator also and solving it would require knowledge of the covariances between the data Y_i and those with the unknown Y.

Consider first the covariance $Cov\{Y_i, Y_j\}$ between the probabilistic log-distances Y_i and Y_j:

$$Cov\{Y_i, Y_j\} = E\{Y_i Y_j\} - E\{Y_i\}E\{Y_j\}$$

In the case of indicator variables, the product $Y_i Y_j$ can take on four values:

$$Y_i Y_j = \begin{cases} y_i^+ y_j^+ & \text{, if } D_i = 1, D_j = 1 \\ y_i^+ y_j^- & \text{, if } D_i = 1, D_j = 0 \\ y_i^- y_j^+ & \text{, if } D_i = 0, D_j = 1 \\ y_i^- y_j^- & \text{, if } D_i = 0, D_j = 0 \end{cases}$$

These four outcomes can be used to compute the above covariance. Some algebraic steps give:

$$Cov\{Y_i, Y_j\} = (y_i^+ - y_i^-)(y_j^+ - y_j^-)Cov\{D_i, D_j\}$$

Similarly, the covariance $Cov\{Y_i, Y\}$ between the data log-distance Y_i and the target log-distance Y is:

$$Cov\{Y_i, Y\} = (y_i^+ - y_i^-)(y^+ - y^-)Cov\{D_i, D\}$$

Observe that this latter covariance $Cov\{Y_i, Y\}$ requires knowledge of y^+ and y^- where y^+ is precisely the unknown log-distance we are trying to evaluate.

Thus, consider instead the correlations expressions $Corr\{Y_i, Y_j\}$ and $Corr\{Y_i, Y\}$:

$$\begin{aligned} Corr\{Y_i, Y_j\} &= \frac{Cov\{Y_i, Y_j\}}{\sqrt{Var\{Y_i\}Var\{Y_j\}}} \\ Corr\{Y_i, Y_j\} &= \frac{(y_i^+ - y_i^-)(y_j^+ - y_j^-)Cov\{D_i, D_j\}}{\sqrt{(y_i^+ - y_i^-)^2 Var\{D_i\}(y_j^+ - y_j^-)^2 Var\{D_j\}}} \\ &= \frac{(y_i^+ - y_i^-)}{|y_i^+ - y_i^-|}\frac{(y_j^+ - y_j^-)}{|y_j^+ - y_j^-|} Corr\{D_i, D_j\} \\ &= Sign[(y_i^+ - y_i^-)(y_j^+ - y_j^-)]Corr\{D_i, D_j\} \\ Corr\{Y_i, Y\} &= Sign[(y_i^+ - y_i^-)(y^+ - y^-)]Corr\{D_i, D\}. \end{aligned}$$

and $Sign[z] = +1$, if $z > 0$ and -1, if $z < 0$.

The log-distance correlations are equal to the data-correlations up to a sign. The latter expression for $Corr\{Y_i, Y\}$ requires knowledge of the $Sign$ term which is likely easier to evaluate than the actual difference $(y_i^+ - y_i^-)$.

Using these correlations, the RV interpretation of the log-distances allows for an evaluation of the tau weights very similar to solving a kriging system. It remains to be seen how these kriging-type weights relate to the exact weights in expression (4).

Interpreting the sign of correlation

Consider the product $\chi_{ij} = Sign[(y_i^+ - y_i^-)(y_j^+ - y_j^-)]$.

$y_i^+ > y_i^-$ means that $D_i = 1$ results in a lesser probability of A occurring than $D_i = 0$, i.e., A is more likely to occur given $D_i = 0$.

Therefore, $\chi_{ij} = 1$ implies that D_i informs A in the same direction as D_j.

Consider now $\chi_i = Sign[(y_i^+ - y_i^-)(y^+ - y^-)]$.

Applying the same reasoning as above, $\chi_i = 1$ implies that $D_i = 1$ informs in the same direction as $D = 1$ (individual datum agreeing with the joint datum).

5 Discussion and conclusions

The purpose of this paper is to establish the tau model as an exact expression for combining probabilistic information. It is shown that the tau weights can be computed for any joint distribution of the data and the unknown. These weights are dependent on the specific sequence of data conditioning. Any sequence is fine as long as the weights for that sequence can be estimated. For any given problem, the most suitable sequence would depend on the specific requirements of the problem. It might be suitable in some cases to condition first to the easiest datum and proceeding to the most difficult datum to be conditioned to. As an example, consider the problem of conditioning geostatistical realizations to point data from well-logs, soft data from seismic-based sources and flow-based data from well-tests. In such case, it would be better to condition first to the point data, then to the seismic information and finally to the flow-based information.

Example of data from different supports

The accompanying paper in this volume (Krishnan, 2004) considers an example of complex data from different supports. The unknown A is the multiple-point rectilinear connectivity function at a particular support of interest. The data D_i are strings of connected high values from smaller and larger supports. The exact tau weights are computed for this example and the effect of data redundancy on these weights are explained. The proposed calibration technique is then used to approximate these weights.

One important result relates to the impact of the common assumption of conditional independence. As noted in this paper, such an assumption leads to a constant tau weight of 1 for all data. It is shown that putting all weights equal to 1 leads to a severe bias of overestimating the target conditional probability. In that case, the various data all compound towards a probability of 1 leading to extremely high probabilities of A occurring. Accounting for data redundancy through the exact expression (4) or through the calibration expression (7) removes that bias.

Future work

The concepts outlined in this paper open many avenues for research:

- The exact tau weights in expression (4) are dependent on the data value. One idea is to approximate these exact tau weights by values which are approximately invariant of the data values. Such tau weights could depend on some homoscedastic measure of data redundancy, eg. covariance or mutual information measure (McEliece, 2002). Analytical measures of data redundancy can help in approximating the tau weights for any given distribution of the data. One would have to evaluate the approximation of the tau weights being independent of the data values. Our conjecture is that such assumption would be much less severe than the commonplace assumption of all tau weights equal to 1 which results from conditional independence.

- The proposed calibration method has several limitations. It does not reflect the generality of the tau model. Typically, the more generality is needed, the more difficult is the inference of the required data statistics. One immediate extension of the proposed method is to allow the weights to go beyond the interval $[0, 1]$.

- The random variable-based approach was developed here only for indicator variables. Both the distance and log distance are non-linear transformations of the conditional probability $P(A|D_i)$. As shown for the indicator case, the correlation between the log-distances is equal to the original indicator data correlation up to a sign. One would also expect convenient analytical expressions if the log-distances are assumed to be Gaussian distributed (Journel, 1980).

References

Bordley, R. F., (1982), *A multiplicative formula for aggregating probability assessments*, Management Science, V. 28, no. 10, pp. 1137-1148.

Genest, C. and Zidek, J. V., (1986), *Combining probability distributions: A critique and an annotated bibliography*, Statistical Science, Vol. 1, pp. 114-135.

Journel, A. G., 1980, *The Lognormal approach to predicting local distributions of selective mining unit grades*, Math. geol., v. 12, No. 4, pgs. 285-303.

Journel, A. G., 2002, *Combining knowledge from diverse sources: An alternative to traditional conditional independence hypothesis*, Math. Geol., Vol. 34, No. 5, 573-596.

Krishnan, S., 2004, *Experimental study of multiple-support, multiple-point dependence and its modeling*, In Proceedings of Geostatistics Congress 2004, Banff, Canada (in this volume).

Liu, Y., 2002, *Improving reproduction of large scale continuity in sequential simulation using structured path*, SCRF Report no. 15, Stanford University.

McEliece, R. J., 2002, *The theory of information and coding*, Cambridge University Press.

AN INFORMATION CONTENT MEASURE USING MULTIPLE-POINT STATISTICS

YUHONG LIU
ExxonMobil Upstream Research Company, P.O.BOX 2189, Houston, TX 77252, USA. E-mail: yuhong@pangea.stanford.edu

Abstract. Multiple-point geostatistics aims at reproducing complex patterns involving many locations at a time, which is much beyond the reach of a two-point variogram model as in traditional geostatistics. In multiple-point geostatistics, sometimes it is necessary to have a quantitative measurement of how informative a data event is with regard to the unknown node, the multiple-point equivalence of a kriging variance. It should be a statistic accounting not only for various possible data configuration and specific data values, but also for the spatial structural information provided by prior geological knowledge. In this paper, we propose two alternative definitions of information content for a multiple-point data event. One is defined as a linear function of the conditional probability, and the other uses entropy for the definition. This information content measure can be widely applied in many occasions in multiple-point simulation. Three applications are presented in the paper. First it is used to rank all unknown nodes to generate a structured path for sequential simulation. Second it is used to decide how to reduce a data event when not enough replicates of it can be found in the training image. Finally it is used to adjust the relative contributions of different data sources in a data integration algorithm. All these applications show an improvement of simulation due to the utilization of this newly defined multiple-point statistic.

1 Introduction

Multiple-point geostatistics (Journel, 1992; Srivastava, 1993; Strebelle, 2000) aims at reproducing complex statistics involving many locations at a time. In multiple-point simulation, at any unsampled node, all conditioning data within its neighborhood are considered as one single data event; then a probability conditioning to this multiple-point data event is derived. This allows capturing pattern information from a training image, which is much beyond the reach of a mere variogram model as in traditional two-point geostatistics.

For various purposes, it is helpful to have a quantitative measurement of how informative a multiple-point data event is with regard to the unknown event to be estimated or simulated. In another word, it is necessary to quantitatively evaluate how much additional information a multiple-point data event brings to the unknown event.

Intuitively, an information content measure should depend on the following two factors:

1. The multiple-point data event, which involves different aspects:

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 1047–1056.

- number of individual nodes in the data event
- configuration of the data event
- specific values of each individual node in the data event

2. Any prior knowledge about the spatial patterns of the variable being simulated

The impact of the first factor is obvious. Figure 1 shows the impact of the second factor. Here *A* denotes the event to be informed at the unsampled node, say, that central node belonging to a certain facies; *B* denotes the multiple-point hard conditioning data event surrounding that node, which includes original sample data and previously simulated values; *P(A|B)* denotes the probability of *A* happening given the conditioning data event *B*. Depending on the training image used, the same B data event can be either very informative or not informative at all.

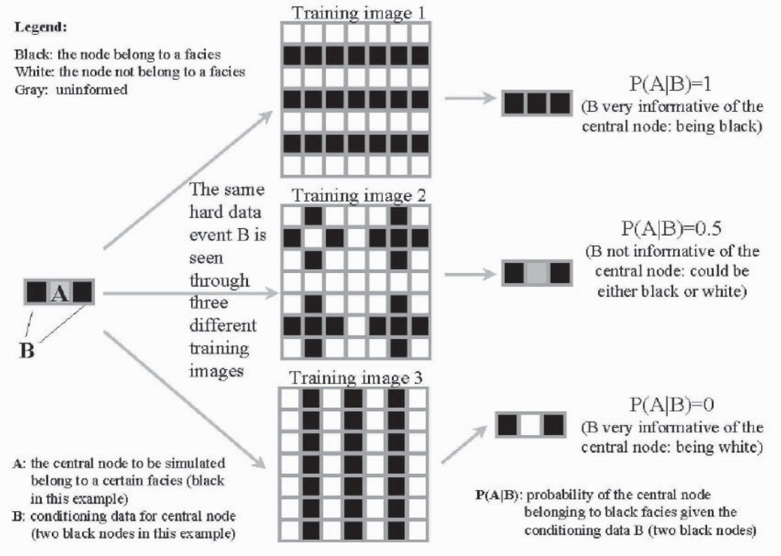

Legend:
Black: the node belong to a facies
White: the node not belong to a facies
Gray: uninformed
Training image 1
P(A|B)=1
(B very informative of the central node: being black)
The same hard data event B is seen through three different training images
Training image 2
P(A|B)=0.5
(B not informative of the central node: could be either black or white)
A
B
Training image 3
P(A|B)=0
(B very informative of the central node: being white)
A: the central node to be simulated belong to a certain facies (black in this example)
B: conditioning data for central node (two black nodes in this example)
P(A|B): probability of the central node belonging to black facies given the conditioning data B (two black nodes)

Figure 1. The same conditioning data event can be either very informative or not informative, depending on the training image used.

We propose to define information content as a function of the conditional probability *P(A|B)*, which can be derived either from solving some kriging/cokriging systems in two-point geostatistics, or from scanning a training image in multiple-point geostatistics. The reason for this definition is that the conditional probability accounts for not only the configuration and values of the multiple-point data event *B*, but also for the prior geological knowledge carried by either the variogram model or the training image. In the following, two alternative definitions are proposed, then the validity of these definitions is shown by their applications to three different occasions.

2 Information content

Intuitively, the information content of a multiple-point data event *B* (denoted as ω_B) with regard to *A* is related to the conditional probability *P(A|B)*. Say, if *P(A|B)=1*, it is certain that *A* is going to happen given the fact that *B* happens, therefore, *B* is very informative of *A*. Similarly for the case when *P(A|B)=0*: it is certain that *A* is NOT going to happen given the fact that *B* happens, hence *B* is again very informative of *A*. In these two cases, the information content of *B* with regard to *A* reaches the maximum. Conversely, if *P(A|B)=0.5*, it is not certain whether *A* is going to happen or not, hence the information content of *B* reaches the minimum. Based on this intuition, we can make a generalized definition of information content as a function of *P(A|B)* satisfying the following conditions:

1. $\omega_B \in [0,1]$;
2. $\omega_B \to 1$ when B is most informative, i.e., when $P(A|B) \to 0$ *or* 1;
3. $\omega_B \to 0$ when B is not informative, i.e., when $P(A|B) \to 0.5$;
4. ω_B decreases monotically within the range *[0,0.5]*;
5. ω_B increases monotically within the range *[0.5,1]*.

For example, all three curves shown in Figure 2 are valid definitions for information

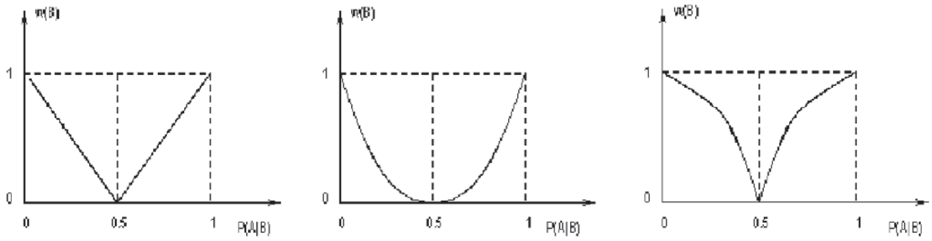


Figure 2. Three possible definitions of information content.

For the case when there is some prior information about A, say, its marginal probability *P(A)*, then the lowest information content should be shifted to the point where *P(A|B)=P(A)*. That is, it measures the "additional" amount of information brought in by data event B besides that provided by the marginal *P(A)*.

In the following, two alternative definitions for the information content are proposed and discussed: first a linear definiation and then an entropy-based definition.

2.1 A LINEAR DEFINTION

For any given multiple-point data event *B* and corresponding conditional probability *P(A|B)*, the information content of *B* with regard to *A*, denoted ω_B, can be defined as a linear function of *P(A|B)*:

$$\omega_B = \begin{cases} \dfrac{P(A\mid B)-P(A)}{1-P(A)} & \text{if } P(A\mid B) \geq P(A) \\ \dfrac{P(A)-P(A\mid B)}{P(A)} & \text{if } P(A\mid B) < P(A) \end{cases} \tag{1}$$

where *P(A)* is the prior probability of the event *A* to be informed, "prior" in the sense that this is what is known about *A* prior to collecting the data event *B*.
The first image of Figure 2 illustrates this definition when *P(A)=0.5*.

Note that this ω_B considers not only the data configuration and values, but also the prior geological knowledge carried by either the variogram model or the training image. This is because *P(A|B)* is derived either from solving a kriging system, or from scanning a training image. Figure 3 illustrates this point: (a) is a training image that captures the curvilinear patterns of channels; (b) shows conditioning well data; (c) shows the average (E-type estimate) of 100 simulated realizations with a multiple-point program, *snesim* (Strebelle, 2000). This E-type map is essentially a *P(A|B)* map. (d) shows the information content ω_B map derived from the *P(A|B)* map using Eq.1. (e) shows the posterior variance of the 100 realizations. It is observed from Figures 3 (d-e) that both the information content and the posterior variance captures the channels' curvilinear structural information displayed by the training image. However, the former can be calculated a prior to simulation while the latter can be calculated only after simulation in multiple-point simulation. Note that at well locations the information content is always the highest, yet the size and shape of their "impact" areas (the high information content area around wells) varies depending on the specific values at well locations and their interaction with other neighboring data according to the training image patterns.

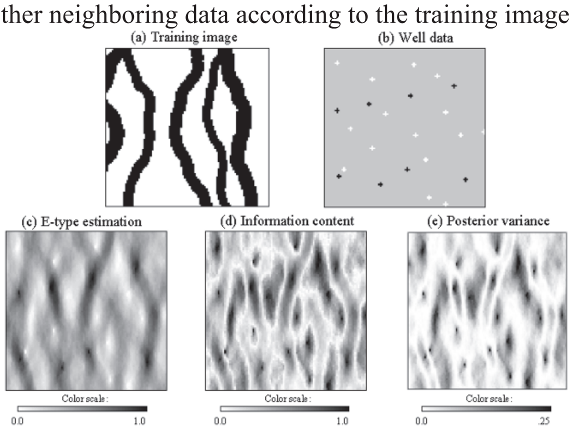


Figure 3. (a) training image; (b) conditional well data; (c) P(A|B) (E-type estimation) map; (d) information content map.

2.2 AN ENTROPY-RELATED DEFINITION

An interesting link is found between entropy and the previously defined information content. Entropy is a core concept in Information Theory (Cover, 1991). An entropy measures the amount of "randomness" of a random variable, say *X*. The entropy of *X*, denoted as *H(X)*, reaches its maximum value when *X* is uniformly distributed, corresponding to minimum information; it reaches the minimum value when there is no uncertainty about *X*, i.e., *X* happens with probability *1* or *0*. The entropy *H(X)* of a discrete random variable *X* with outcomes *x* and probability *p(x)* is defined as:

$$H(X) = -\sum_{\text{all possible } x} p(x)\log[p(x)]$$

The log is generally to the base *2*. *0log0* is taken to be *0*.

For a binary categorical variable, say, presence or absence of a certain facies event *A* at an unknown node, each with probability *p* and *1-p*, the entropy *H* is:

$$H = -p\log(p) - (1-p)\log(1-p) \qquad (2)$$

Figure 4a shows entropy as a function of *p* for a binary categorical random variable. It is a concave function of *p*: with maximum *1* when *p=0.5* and minimum *0* when *p=0* or *1*.

If Figure 4a is flipped upside down (see Figure 4b), we get a valid measure of information content ω, defined as:

$$\omega = 1 - H = 1 + p\log(p) + (1-p)\log(1-p) \qquad (3)$$

This expression is consistent with the previous requirements for an information content measure: when p approaches *0* or *1*, the information content ω monotonically increases to its highest value *1*; conversely, when *p* approaches *0.5*, the information content ω monotonically decreases to its lowest value *0* (compare Figure 4b with Figure 2b).

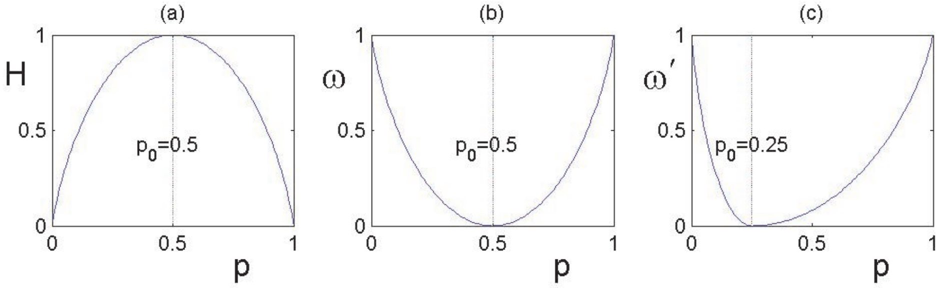


Figure 4. (a) entropy; (b) information content before standardized by the marginal probability; (c) information content after standardized by the marginal probability.

Eq.3 is actually the relative entropy between the uniform distribution (the least informative distribution) and the probabilistic distribution of the random variable A. In information theory, this relative entropy or Kullback Leibler distance (Cover, 1992) is used as a measure of the distance between two probabilistic distributions. In statistics, relative entropy *D(p||q)* arises as the expected logarithm of the likelihood ratio, defined as a measure of the inefficiency of assuming that the distribution is *q* when it is actually *p*. It can be thought as a measure of the "distance" between two probability distributions *p* and *q*. It is defined as,

$$D(p \parallel q) = \sum_x p(x)\log\frac{p(x)}{q(x)} = E_p\left\{\log\frac{p(x)}{q(x)}\right\}$$

Where $E_p\{\cdot\}$ is the expected value taken over the probability distribution *p*.

It can be shown (Cover, 1991) that, for an *m*-category random variable *X*, the relative entropy of any distribution *p* versus a uniform distribution *u* is:

$$D(p \parallel u) = \sum_x p(x)\log\frac{p(x)}{u(x)} = \log_2(m) - H(x) \qquad (4)$$

For binary cases, *m=2*, the above equation becomes:

$$D(p \parallel u) = \log_2(m) - H(x) = 1 - H = 1 + p\log(p) + (1-p)\log(1-p)$$

This is the information content defined in Eq.3. Therefore Eq.3 can be seen as the distance of a probability distribution to the uniform distribution, i.e., the least informative distribution. The information content is the lowest when *p* is the same as the uniform distribution, and it increases when *p* is farther away from that uniform distribution.

A relative entropy is always non-negative and is zero if and only if the two distributions are exactly the same. Note that, however, this relative entropy is not a true distance because it is not symmetric and does not satisfy the triangle inequality, and *D(p||q)* is not equal to *D(q||p)* in general.

Note that the information content measure defined by Eq. 4 has the advantage over the one defined by Eq. 3 or the linear definition defined by Eq. 1, because it allows multiple (>2) categories or even continuous variable.

Eq.3 applies to cases when no prior information is available, or the prior probability $p_0=0.5$. In practice, some prior information may be available, and the prior probability p_0 could be different from 0.5. The definition Eq.3 is then modified to yield the lowest value at $p=p_0\neq 0.5$ (see Figure 4c):

$$\omega' = 1 - H' = 1 + p'\log(p') + (1-p')\log(1-p') \qquad (5)$$

Where p' is the probability standardized by the prior probability p_0, defined as:

$$p' = \begin{cases} \dfrac{p}{2p_0} & \text{if } p \le p_0 \\[2ex] \dfrac{p - 2p_0 + 1}{2(1-p_0)} & \text{if } p > p_0 \end{cases} \qquad (6)$$

Using Eqs. 5 and 6, along with the knowledge of the marginal probability *P(A)* and the conditional probability *P(A|B)*, we can define the information content of *B* data event about *A* as:

$$\omega_B = 1 + P'(A\,|\,B)\log(P'(A\,|\,B)) + (1 - P'(A\,|\,B))\log(1 - P'(A\,|\,B)) \qquad (7)$$

Where *P'(A|B)* is the probability standardized by the marginal probability *P(A)*:

$$P'(A\,|\,B) = \begin{cases} \dfrac{P(A\,|\,B)}{2P(A)} & \text{if } \mathrm{P(A\ |\ B)} \le P(A) \\[2ex] \dfrac{P(A\,|\,B) - 2P(A) + 1}{2(1-P(A))} & \text{if } P(A\,|\,B) > P(A) \end{cases} \qquad (8)$$

3 Application Case Studies

The information content concept can be widely applied to many different occasions. Only three applications are discussed in details in this section. Other applications include using information content to adjust the amount of probability perturbation during servosystem correction (to honor a target global proportion) in *snesim* program, etc. They are not discussed in this paper due to limited paper length.

3.1 STRUCTURED VISITING PATH

In sequential simulation, multiple equi-probable realizations can be generated through changing the random visiting path. One problem with a purely random path concept is that a nodal value may be drawn with few or no conditioning data, simply because it is visited too early. This problem becomes prominent with the multiple-point simulation algorithm, *snesim* program (Strebelle, 2000). Specifically, an accidentally simulated value with low probability could propagate to its immediate neighboring nodes and forbids large-scale continuity with other values simulated far away. The resulting realizations would then display discontinuities of long-range structures such as channels failing to cross the entire field. Figure 5 shows the simulation proceeding of one such realization. Image 1 shows the original well data. Image 8 shows the final realization.

The other images show the intermediate steps when only part of the field is simulated. It is found that starting from Image 3, a mud node (in the small square) is simulated due to limited conditioning data, then this simulated mud node gets propagated around it in subsequent simulation, finally resulting in discontinuity of the simulated channel.

To address this problem, Liu (2002) proposed to simulate along a structured path: visit first the better informed nodes, then proceed to the less informed nodes. The information content is used to rank the nodes for their visiting order. Note that the two definitions by Eq. 1 or 5 yield the same results because their ranking orders are the same. This information content-based structured path accounts for much more information than spiral away from sample data, specifically:

- it considers not only the sample data, but also the **previously simulated values**
- it considers not only the location of any single datum relateive the the location of the unsimulated node, but also the **multiple-point data configuration**
- it considers both data locations and **data values**
- it accounts for the **prior geological knowledge**

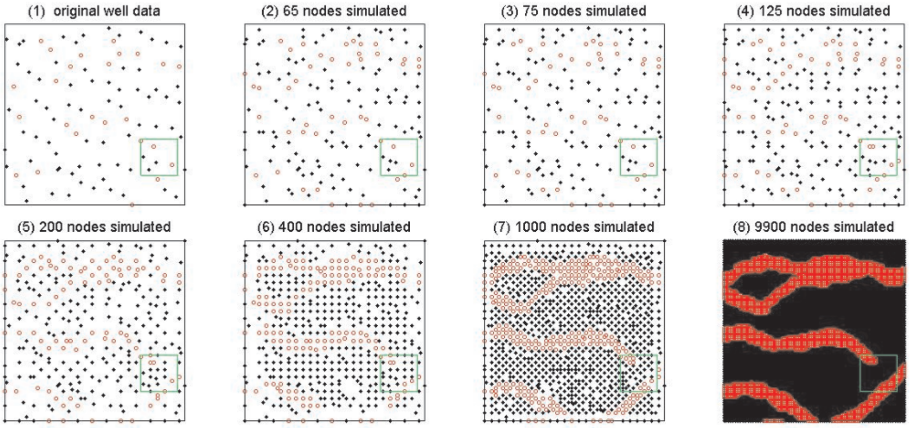


Figure 5. The simulation proceeding of one *snesim* realization. Circles represent channel and small dots represent mud nodes.

The original *snesim* program is modified to implement this structured path, along with a better inference of *P(A|B)* (see the following subsection). Figures 6a-c shows the training image, reference field, and conditioning well data. Figures 6d-g show two realizations respectively by the original and modified *snesim* program. It is observed that the long-range channels are better reproduced by the modified *snesim* program.

3.2 IMPROVED INFERENCE OF *P(A|B)*

When simulating with *snesim* program, at any unsampled node, the data found in its neighborhood constitute a conditioning data event *B*. Prior to simulation, a training image is scanned and all replicate numbers of different data events found are stored in a search tree. Then in simulation, the number of training replicates of a specific data event *B* is retrieved from the search tree. The training proportion of the central node belonging to a certain facies category is then taken as the conditional probability *P(A|B)*.

To avoid unreliable inference of that probability, the total number of replicates is required to be no less than an input minimum value. When there are not enough replicates for the data event *B*, that data event has to be reduced by dropping conditioning data one by one until enough replicates can be found from the search tree. The original *snesim* program reduces this data event *B* by dropping the furthest away data. The decision of dropping the furthest away data amounts to value closer data more than further away data, even though the latter carries information about long-range structures. As a result, large-scale structures such as continuous channels may get broken during simulation.

Instead of dropping the furthest away data, it is suggested to drop the data that are less "certain", that is, those nodes with lower information content at the time of their simulation. Specifically, during a multi-grid simulation, we record the conditional probability *P(A|B)* from which each simulated node has been drawn during simulation at the previous coarser grid. The information content ω_B proposed in Eq. 1 is calculated from this recorded *P(A|B)* value. This ω_B is set to 1 if that node identifies an original sample datum. When not enough replicates of *B* are found, instead of dropping the furthest away data, we now drop the data with the lowest $\omega_{B,}$ calculated from the recorded *P(A|B)*. Through this modification, large-scale structural information provided by farther away data is kept when *B* data event need be reduced. Compare Figures 6(d-e) with 6(f-g), it is observed that this new dropping scheme, associated with a structured path, helps to alleviate the channel discontinuity problem when simulating channel with the *snesim* program.

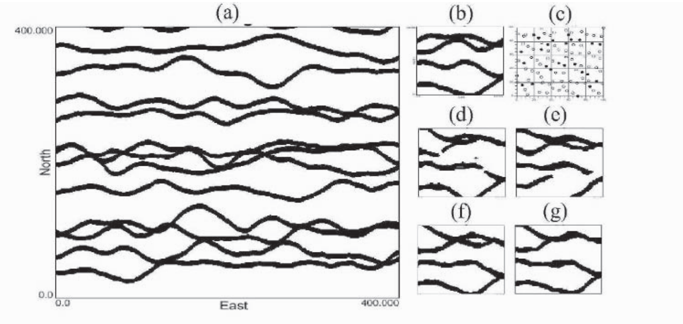


Figure 6. (a) Training image; (b) reference facies field; (c) conditioning well data; (d-e) two realizations by original *snesim*; (f-g) two realizations by modified *snesim* program.

3.3 DATA INTEGRATION

In multiple-point simulation, one way to integrate data of different sources (e.g., hard data *B* and soft data *C*) is to first obtain the individual probabilities *P(A|B)* and *P(A|C)*, each conditioned to a single data source *B* and *C*. Then combine them into an updated probability conditioned to all data sources: *P(A|B,C)*. Journel (2002) proposed a "Permanence of Updating Ratios" paradigm to accomplish this. The basic assumption of this algorithm is that the relative contribution of data event C is the same before and after knowing B:

$$\frac{x}{b} = \frac{c}{a} \tag{9}$$

where, a, b, c and x represent distances (≥ 0) to the event A occurring defined as:

$$a = \frac{1-P(A)}{P(A)}, \quad b = \frac{1-P(A\,|\,B)}{P(A\,|\,B)}, \quad c = \frac{1-P(A\,|\,C)}{P(A\,|\,C)}, \quad x = \frac{1-P(A\,|\,B,C)}{P(A\,|\,B,C)}$$

Zhang and Journel (2003) later showed that this approach is equivalent to a Bayesian updating under conditional independence of B and C given A. To account for dependence between B and C data, they proposed a generalization using a power model involving two parameters, τ_B and τ_C (Eq. 10):

$$\frac{x}{a} = \left(\frac{b}{a}\right)^{\tau_B} \left(\frac{c}{a}\right)^{\tau_C} \tag{10}$$

Note that this generalized Eq. 10 is equal to Eq. 9 when $\tau_B = \tau_C = 1$.

The challenge now is to determine the parameter values τ_B and τ_C (called τ-model hereafter), which should depend on the two data events B and C taken simultaneously. One method, although imperfect because it does not account explicitly for the B, C data dependence, is to link τ_B and τ_C to the information content of B and C:

$$\tau_B = \frac{\omega_B}{\omega_B + \omega_C}, \qquad \tau_C = \frac{\omega_C}{\omega_B + \omega_C} \tag{11}$$

where ω_B and ω_C are the measures of information content of the two multiple-point conditioning data events B and C.

The idea of Eq. 11 is to tune up or down the impact of B or C corresponding to their respective information content: if B is very informative while C is not, τ_B is tuned up and τ_C is tuned down, and vice versa.

A synthetic data set is generated to test this approach (Liu, 2002). Based on the same facies reference field A, three sets of data, each of a different quality, are generated respectively for hard data B and soft data C. They form 9 different (B, C) data combinations. For each combination, we can obtain $P(A|B)$, $P(A|C)$ and $P(A|B,C)$. $P(A|B,C)$ is assumed to be unknown and taken as the reference. Then for each (B,C) combination, we combine the individual probability fields $P(A|B)$ and $P(A|C)$ into an estimated $P^*(A|B,C)$ using one of the following τ-models:

- τ-model 1: $\tau_B = \tau_C = 1$, which is equivalent to the original permanence of updating ratio algorithm (Eq. 9), that is, B and C are conditionally independent given A.
- τ-model 2: Use Eq. 11 to calculate τ_B and $\tau_{C,}$ where the information content ω_B and ω_C are calculated from the linear definition (Eq. 1).
- τ-model 3: Similar to τ-model 2,but use entropy-related information content (Eq.7).

The estimated $P^*(A|B,C)$ field by each τ-model is compared with the reference $P(A|B,C)$ field to obtain the mean squared error (MSE) of the estimated $P^*(A|B,C)$. A smaller MSE indicates a better estimation. This comparison is done for each of the nine (B, C) data combinations.

Table 1 shows the MSE by the three different τ-models for the nine different (B,C) data combinations. It is observed that τ-models 2 and 3 always yield smaller MSE than τ-model 1, that is, utilizing information content helps to reduce the estimation error.

Another observation is that τ-models 2 and 3 yield similar results, that is, the linear and entropy-related information content are equivalent. However, the latter is preferred because of two reasons: (1) The linear definition can deal with only the binary categorical cases, while the entropy-related definition can be extended to the cases of multi-category (through Eq. 4) or even continuous variable. (2) The entropy-related definition has the potential to be further extended to account for data dependence and redundancy through utilizing some core concepts in Information Theory, for example, mutual information, chain rule of data communication, etc.

B data	good			fair				poor	
C data	good	fair	poor	good	fair	poor	good	fair	poor
Model 1	.075	.077	.071	.092	.179	.174	.080	.169	.216
Model 2	.059	.067	.068	.075	.153	.165	.077	.168	.211
Model 3	.059	.067	.068	.076	.152	.165	.077	.167	.211

Table 1. MSE of estimated *P*(A|B,C)* versus reference *P(A|B,C)*.

4 Conclusions

A measure of information content for a multiple-point data event is proposed. It is a multiple-point statistic measuring how much additional information is brought in by a data event to the unknown event. Two alternative definitions of information content are proposed, both are functions of a conditional probability: a linear definition and an entropy-related definition. They provide similar results, but the latter is preferred due to its potential to be further extended. The validity of these two definitions is shown by three application case studies.

References

Cover, T. and Thomas, J.,: Elements of Information Theory, John Wiley, New York, 576p, 1991.

Deutsch, C.V. and Journel, A.G., GSLIB: Geostatistical Software Library and User's Guide, Oxford University Press, 1992.

Deutsch, C. and Wang, L.: Hierarchical Object-Based Stochastic Modeling of Fluvial Reservoirs. Mathematical Geology, vol. 28, no. 7, 1996, p 857-880.

Gilbert, R., Liu, Y., Abriel, W. and Preece, R.: Reservoir Modeling Integrating Various Data and at Appropriate Scales, The Leading Edge, Vol.23, no. 8, Aug. 2004, p784-788.

Guardiano, F. and Srivastava, R.M.: "Multivariate Geostatistics: Beyond Bivariate Moments", Geostatistics-Troia, A. Soares (ed.), Kluwer Academic Publications, Dordrecht, 1993, vol 1, p 113-114.

Journel, A: Geostatistics: Roadblocks and Challenges. In A. Soares (Ed.), Geostatistics-Troia, Kluwer Academic Publ., Dordrech, 1992, p 213-224.

Journel, A: Combining knowledge from diverse sources: an alternative to traditional data independence hypotheses, Mathematical Geology, vol. 34, no. 5, 2002.

Liu, Y.: Downscaling Seismic Data into A Geological Sound Numerical Model, Ph.D. dissertation, Department of Geological and Environmental Science, Stanford University, Stanford, 2003, pp. 202.

Liu, Y., Harding, A., Abriel, W. and Strebelle, S.: Multiple-Point Simulation Integrating Wells, 3D Seismic Data and Geology, AAPG Bulletin, Vol. 88, No.7, 2004.

Liu, Y., Journel, A.: Improving Simulation with a Structured Path Guided by Information Content, Mathematical Geology, vol. 36, no. 8, 2004.

Liu, Y.: Data Integration in Multiple-point Simulation, Mathematical Geology, manuscript submitted.

Strebelle, S.: Sequential Simulation Drawing Structures from Training Images, Ph.D. dissertation, Department of Geological and Environmental Sciences, Stanford University, Stanford, 2000.

Strebelle, S., Payrazyan, K. and Caers, J.: Modeling of a Deepwater Turbidite Reservoir Conditional to Seismic Data Using Multiple-Point Geostatistics, SPE 77425, SPE Annual Technical Conference and Exhibition, San Antonio, Texas, 2002.

Zhang, T. and Journel, A. G.: Merging Prior Geological Structure and Local Data: the mp Geostatistics Answer, SCRF Annual Report No.16, vol. 2, 2003.

INTERNAL CONSISTENCY AND INFERENCE OF CHANGE-OF-SUPPORT ISOFACTORIAL MODELS

XAVIER EMERY and JULIÁN M. ORTIZ
Department of Mining Engineering, University of Chile
Avenida Tupper 2069, Santiago, Chile, 837 0451

Abstract. Bivariate isofactorial models are used for global or local change-of-support applications. However, so far, their variogram analysis is complicated and may lead to mathematical inconsistencies. In this paper, we propose an alternative approach for internally consistent variogram inference, which consists in deriving the simple and cross variograms at point and block supports from the variogram of transformed data, by randomizing the sample locations within the blocks (regularization). This approach is illustrated with the discrete Hermitian model, for which we provide guidelines for parameter inference and emphasize the limitations of the extreme cases: discrete Gaussian and mosaic models. A case study is presented with an application of the Hermitian model to a mining dataset, which consists of drillhole samples measuring the grade in a porphyry copper deposit.

1 Introduction

Change of support is a key problem in application fields such as ore reserve estimation, environmental and soil sciences. A full answer to this issue requires specifying the spatial distribution of the random function that describes the regionalized variable under study. Usually, it cannot be handled analytically and conditional simulations are used, which is time-consuming. Another solution is provided by parametric models such as isofactorial models that rely on bivariate distributions. In these models, the space is divided into non-overlapping blocks which are identical up to a translation, and the sample locations are randomized within the blocks (Matheron, 1976b, p. 243; Chilès and Delfiner, 1999, p. 439). In order to avoid confusion henceforth, the bold character $\mathbf{x}$ will refer to a *fixed* sample location, whereas $\underline{\mathbf{x}}$ (underlined) to a *random* location. If several samples belong to the same block, their random locations are assumed to be independent inside this block.

Two types of isofactorial models must be distinguished, depending on whether a global or a local description is needed.

2 Global isofactorial models

The global change-of-support models are based on the following assumptions:

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 1057–1066.

- The point-support and block-support variables, denoted by $Z_{\mathbf{x}}$ and Z_v hereafter, can be transformed into other variables (denoted by $Y_{\mathbf{x}}$ and Y_v respectively), on which the isofactorial properties will be stated:

  $Z_{\mathbf{x}} = \phi(Y_{\mathbf{x}})$ and $Z_v = \phi_v(Y_v)$ where ϕ and ϕ_v are the *transformation functions*.

- For any block v and any random location $\underline{\mathbf{x}}$ uniformly distributed inside v, the pair $\{Y_{\mathbf{x}},Y_v\}$ has an asymmetric isofactorial distribution (Matheron, 1984b, p. 451):

  $$\forall \underline{\mathbf{x}} \in v, \forall y, y' \in \mathsf{R}, \mathrm{f}_{\mathbf{x}v}(y,y') = \mathrm{f}_{\mathbf{x}}(y)\mathrm{f}_v(y')\{1 + \textstyle\sum_{p\geq 1} \mathrm{T}_p(\underline{\mathbf{x}},v)\chi_p^{\mathbf{x}}(y)\chi_p^v(y')\} \qquad (1)$$

  where $\mathrm{f}_{\mathbf{x}}$ and f_v are the marginal *pdf* of $Y_{\mathbf{x}}$ and Y_v, $\{\mathrm{T}_p(\underline{\mathbf{x}},v), p \in \mathsf{N}^*\}$ is a set of real coefficients, whereas $\{\chi_p{}^{\mathbf{x}}, p \in \mathsf{N}^*\}$ and $\{\chi_p{}^v, p \in \mathsf{N}^*\}$ are orthonormal functions for $\mathrm{L}^2(\mathsf{R},\mathrm{f}_{\mathbf{x}})$ and $\mathrm{L}^2(\mathsf{R},\mathrm{f}_v)$ respectively (they are called the *factors* of the isofactorial model).

Now, the transformation functions at both supports can be expanded into the factors:

$$\forall y \in \mathsf{R}, \phi(y) = \phi_0^{\mathbf{x}} + \textstyle\sum_{p\geq 1} \phi_p^{\mathbf{x}}\, \chi_p^{\mathbf{x}}(y) \;\text{ and }\; \phi_v(y) = \phi_0^v + \textstyle\sum_{p\geq 1} \phi_p^v\, \chi_p^v(y) \qquad (2)$$

Because of Cartier's relation (Matheron, 1984a, p. 425; Chilès and Delfiner, 1999, p. 426) and Equation (1), the coefficients of the previous expansions are linked together:

$$\phi_0^v = \phi_0^{\mathbf{x}} \;\text{ and }\; \forall p \in \mathsf{N}^*,\ \phi_p^v = \mathrm{T}_p(\underline{\mathbf{x}},v)\,\phi_p^{\mathbf{x}} \qquad (3)$$

The global model only requires specifying a set of parameters $\{\mathrm{T}_p(\underline{\mathbf{x}},v), p \in \mathsf{N}^*\}$ so that the joint density between point and block supports (Eq. 1) is always positive and the block-support variance (known after a variogram analysis of $Z_{\mathbf{x}}$) is honored:

$$\mathrm{var}(Z_v) = \textstyle\sum_{p\geq 1} (\phi_p^v)^2 = \textstyle\sum_{p\geq 1} [\mathrm{T}_p(\underline{\mathbf{x}},v)\,\phi_p^{\mathbf{x}}]^2 \qquad (4)$$

3 Local isofactorial models

Local models rely on the stronger assumption that any pair of values from $Y_{\mathbf{x}}$ or Y_v (not only the collocated values like in Eq. 1) follows an isofactorial distribution (Matheron, 1984b, p. 450; Chilès and Delfiner, 1999, p. 443) (Figure 1):

$$\forall (\underline{\mathbf{x}},\underline{\mathbf{x}}'),\ \mathrm{f}_{\mathbf{x}\mathbf{x}'}(y,y') = \mathrm{f}_{\mathbf{x}}(y)\mathrm{f}_{\mathbf{x}}(y')\{1+\textstyle\sum_{p\geq 1} \mathrm{T}_p(\underline{\mathbf{x}},\underline{\mathbf{x}}')\chi_p^{\mathbf{x}}(y)\chi_p^{\mathbf{x}}(y')\} \qquad (5)$$

$$\forall (\underline{\mathbf{x}},v'),\ \mathrm{f}_{\mathbf{x}v'}(y,y') = \mathrm{f}_{\mathbf{x}}(y)\mathrm{f}_v(y')\{1+\textstyle\sum_{p\geq 1} \mathrm{T}_p(\underline{\mathbf{x}},v')\chi_p^{\mathbf{x}}(y)\chi_p^v(y')\} \qquad (6)$$

$$\forall (v,v'),\ \mathrm{f}_{vv'}(y,y') = \mathrm{f}_v(y)\mathrm{f}_v(y')\{1+\textstyle\sum_{p\geq 1} \mathrm{T}_p(v,v')\chi_p^v(y)\chi_p^v(y')\} \qquad (7)$$

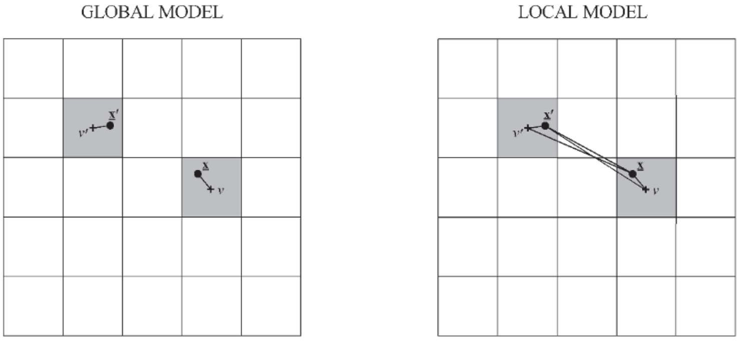


Figure 1. Global and local models: all paired values follow an isofactorial distribution

These distributions imply that two factors with different orders have no spatial cross-correlation and that their simple and cross correlograms are:

$$\forall p \in \mathsf{N}^*, \operatorname{cov}\{\chi_p^{\mathbf{x}}(Y_{\underline{\mathbf{x}}}), \chi_p^{\mathbf{x}}(Y_{\underline{\mathbf{x}}'})\} = \mathrm{T}_p(\underline{\mathbf{x}}, \underline{\mathbf{x}}') \tag{8}$$

$$\forall p \in \mathsf{N}^*, \operatorname{cov}\{\chi_p^{\mathbf{x}}(Y_{\underline{\mathbf{x}}}), \chi_p^{v}(Y_{v'})\} = \mathrm{T}_p(\underline{\mathbf{x}}, v') \tag{9}$$

$$\forall p \in \mathsf{N}^*, \operatorname{cov}\{\chi_p^{v}(Y_{v}), \chi_p^{v}(Y_{v'})\} = \mathrm{T}_p(v, v') \tag{10}$$

Accounting for Equation (2) and for the orthonormality of the factors for the bivariate distributions, the covariances of $Z_{\mathbf{x}}$ and Z_v can be expanded as follows:

$$\operatorname{cov}(Z_{\underline{\mathbf{x}}}, Z_{\underline{\mathbf{x}}'}) = \sum\nolimits_{p\geq 1} (\phi_p^{\mathbf{x}})^2 \operatorname{cov}\{\chi_p^{\mathbf{x}}(Y_{\underline{\mathbf{x}}}), \chi_p^{\mathbf{x}}(Y_{\underline{\mathbf{x}}'})\} = \sum\nolimits_{p\geq 1} (\phi_p^{\mathbf{x}})^2\, \mathrm{T}_p(\underline{\mathbf{x}}, \underline{\mathbf{x}}') \tag{11}$$

$$\operatorname{cov}(Z_{\underline{\mathbf{x}}}, Z_{v'}) = \sum\nolimits_{p\geq 1} \phi_p^{\mathbf{x}}\, \phi_p^{v} \operatorname{cov}\{\chi_p^{\mathbf{x}}(Y_{\underline{\mathbf{x}}}), \chi_p^{v}(Y_{v'})\} = \sum\nolimits_{p\geq 1} (\phi_p^{\mathbf{x}})^2\, \mathrm{T}_p(\underline{\mathbf{x}}, v)\, \mathrm{T}_p(\underline{\mathbf{x}}, v') \tag{12}$$

$$\operatorname{cov}(Z_{v}, Z_{v'}) = \sum\nolimits_{p\geq 1} (\phi_p^{v})^2 \operatorname{cov}\{\chi_p^{v}(Y_{v}), \chi_p^{v}(Y_{v'})\} = \sum\nolimits_{p\geq 1} (\phi_p^{\mathbf{x}})^2\, \mathrm{T}_p^2(\underline{\mathbf{x}}, v)\, \mathrm{T}_p(v, v') \tag{13}$$

Now, because of the randomization of the sample locations within the blocks, all these covariances are equal (except if $\underline{\mathbf{x}} = \underline{\mathbf{x}}'$ in Eq. 11, in which case the sample variance is obtained) (Chilès and Delfiner, 1999, p. 441). A term-to-term identification leads to:

$$\forall p \in \mathsf{N}^*,\ \mathrm{T}_p(\underline{\mathbf{x}}, \underline{\mathbf{x}}') = \left| \begin{array}{l} \mathrm{T}_p(\underline{\mathbf{x}}, v)\, \mathrm{T}_p(\underline{\mathbf{x}}, v') \text{ if } \underline{\mathbf{x}} \neq \underline{\mathbf{x}}' \\ 1 \text{ otherwise} \end{array} \right. \tag{14}$$

$$\forall p \in \mathsf{N}^*,\ \mathrm{T}_p(\underline{\mathbf{x}}, v') = \mathrm{T}_p(\underline{\mathbf{x}}, v)\, \mathrm{T}_p(v, v') \tag{15}$$

These equations amount to a Markov-type hypothesis for the point and block-support variables, similar to the one often used in association with collocated cokriging (Chilès and Delfiner, 1999, p. 305). Given a block-support value, any collocated point-support information is independent of the nearby information (block or sample values), a property also known as *conditional independence* (Matheron, 1984b, p. 451).

4 Proposed approach for variogram inference

Commonly, the factor correlograms $\{\mathrm{T}_p(v,v'), p \in \mathsf{N}^*\}$ are deduced from the first one, $\mathrm{T}_1(v,v')$, by a simple analytical expression. In practice, given a covariance model for $Z_{\mathbf{x}}$ and therefore for Z_v, Equation (13) is inverted to give a discretized approximation of $\mathrm{T}_1(v,v')$ on which a model is fitted (Rivoirard, 1994, p. 90; Chilès and Delfiner, 1999, p. 441). However, this approach leads to three difficulties:

- The structural analysis is intricated: the whole model relies on the block-support correlogram $\mathrm{T}_1(v,v')$ which is not data-charged, i.e. it is obtained indirectly (through the transformation function ϕ_v) and fitted without any block-support data.

- Experience has shown that the discretized correlogram $\mathrm{T}_1(v,v')$ may not be positive definite (Wackernagel, 2003, p. 267), which proves that the variogram of $Z_{\mathbf{x}}$ is not always consistent with the transformation functions or the isofactorial assumptions (Eq. 5 to 7).

- Which correlograms are allowable for the block-support transformed variable Y_v? Indeed, not every model can be used since it refers to a block-support variable: for instance, a pure nugget effect is not conceivable, even in the bigaussian framework.

A better way to perform the variogram analysis would consist of the following steps:

i) Transform the fixed-location variable $Z_{\mathbf{x}}$ into $Y_{\mathbf{x}}$ (classical procedure).

ii) Perform the structural analysis of $Y_{\mathbf{x}}$ and obtain its correlogram $\mathrm{T}_1(\mathbf{x},\mathbf{x}')$.

iii) Compute the correlogram of the random-location variable $Y_{\underline{\mathbf{x}}}$: except for the zero distance, this amounts to regularizing $\mathrm{T}_1(\mathbf{x},\mathbf{x}')$ (Chilès and Delfiner, 1999, p. 79):

$$\forall \underline{\mathbf{x}} \in v, \underline{\mathbf{x}}' \in v', \mathrm{T}_1(\underline{\mathbf{x}},\underline{\mathbf{x}}') = \left| \begin{array}{l} \dfrac{1}{|v|^2} \displaystyle\int_v \int_{v'} \mathrm{T}_1(\mathbf{u},\mathbf{v})\, d\mathbf{u}\, d\mathbf{v} \quad \text{if } \underline{\mathbf{x}} \neq \underline{\mathbf{x}}' \\ 1 \text{ otherwise} \end{array} \right. \tag{16}$$

in which $|v|$ is the volume of block v. By comparing with Equation (14), it follows:

$$\forall \underline{\mathbf{x}} \in v, \mathrm{T}_1(\underline{\mathbf{x}}, v) = \frac{1}{|v|} \sqrt{\int_v \int_v \mathrm{T}_1(\mathbf{u},\mathbf{v})\, d\mathbf{u}\, d\mathbf{v}} \tag{17}$$

This condition is part of a more complete system of equations that can be obtained if the samples are no longer randomized into the blocks (Matheron, 1976b, p. 241; Chilès and Delfiner, 1999, p. 438):

$$\forall p \in \mathsf{N}^*, \forall \mathbf{x} \in v, \mathrm{T}_p(\mathbf{x}, v) = \frac{1}{|v|} \sqrt{\int_v \int_v \mathrm{T}_p(\mathbf{u},\mathbf{v})\, d\mathbf{u}\, d\mathbf{v}} \tag{18}$$

However, such fixed-location models are not used as they are not fully consistent: Equation (18) leads to a non-positive *pdf* for the point-block distribution (Eq. 1).

iv) Deduce the other correlograms $\{T_p(\underline{\mathbf{x}},\underline{\mathbf{x}}'), p \geq 2\}$, $\{T_p(\underline{\mathbf{x}},v'), p \in \mathsf{N}^*\}$ and $\{T_p(v,v'), p \in \mathsf{N}^*\}$, considering the isofactorial model at hand and Equations (14) and (15).

v) Make sure that all these correlograms lead to positive bivariate *pdf* (Eq. 5 to 7). Otherwise, go back to step ii) and choose a different model for $T_1(\mathbf{x},\mathbf{x}')$.

vi) Validate the model: a simple way is to fit and regularize the covariance of $Z_\mathbf{x}$ and compare it to the function obtained by Equation (13). Now, the structural analysis of $Z_\mathbf{x}$ should be used only for validation purposes, not for deriving the correlograms of $Y_\mathbf{x}$ or Y_v, as in the traditional approach. In this respect, Chilès and Delfiner (1999, p. 442) propose the following formula:

$$\operatorname{cov}(Z_v, Z_{v'}) = \frac{1}{|v|^2} \int_v \int_{v'} \sum_{p\geq 1} (\phi_p^{\mathbf{x}})^2 \, T_p(\underline{\mathbf{x}},\underline{\mathbf{x}}')\, d\mathbf{x}\, d\mathbf{x}' \tag{19}$$

but such relation is an approximation: if T_p refers to the random-location variable ($Y_{\underline{\mathbf{x}}}$), the correct formula is given by Equation (13); conversely, the fixed-location variable ($Y_\mathbf{x}$) is not necessarily isofactorial, hence formula (19) is not valid if T_p refers to this variable. Another helpful tool for validating an isofactorial model is the analysis of the variograms of order less than 2, as detailed in the next section.

In the proposed approach, the structural model is obtained directly after the *transformed* data ($Y_\mathbf{x}$), which are those used in the applications of isofactorial models, e.g. uniform conditioning, disjunctive kriging, conditional expectation or conditional simulations (Guibal and Remacre, 1984; Hu, 1988; Rivoirard, 1994; Chilès and Delfiner, 1999, p. 445-448 & p. 573; Emery, 2002, p. 96; Emery and Soto Torres, 2005).

5 Discrete Hermitian model

5.1 GENERAL PRESENTATION

In this model, the marginal distributions are standard Gaussian and the factors for both the point and block supports are the normalized Hermite polynomials. The general form for the factor cross-correlograms (Eq. 9) is (Matheron, 1976a, p. 230):

$$\forall p \in \mathsf{N}^*, \forall \underline{\mathbf{x}}, v', \; T_p(\underline{\mathbf{x}}, v') = E[R(\underline{\mathbf{x}}, v')^p] \tag{20}$$

where $R(\underline{\mathbf{x}},v')$ is a random variable that lies in [-1,1] and depends on the locations of $\underline{\mathbf{x}}$ and v' (or, in the stationary framework, on their separation vector only). Under this condition, the point-support factor correlograms are defined by Equation (14):

$$\forall p \in \mathsf{N}^*, \forall \underline{\mathbf{x}} \neq \underline{\mathbf{x}}', \; T_p(\underline{\mathbf{x}}, \underline{\mathbf{x}}') = E[R(\underline{\mathbf{x}}, v)^p]\, E[R(\underline{\mathbf{x}}, v')^p] \tag{21}$$

The generic term of this succession appears as the product of the moments of order p of two random variables. It is therefore identified with the moment of order p of a random variable $R(\underline{\mathbf{x}},\underline{\mathbf{x}}')$ which is the product of two independent random variables with the same distributions as $R(\underline{\mathbf{x}},v)$ and $R(\underline{\mathbf{x}},v')$. This statement guarantees the positivity of the bivariate point-support distribution (Eq. 5) (Matheron, 1976a, p. 230). However, so far such model with random location samples has hardly been used. A simple example is presented hereafter and a guideline is proposed to infer and validate the parameters.

5.2 AN EXAMPLE: THE BETA MODEL

Suppose that $R(\underline{\mathbf{x}},v')$ follows a beta distribution with parameters $\{\beta\ r_{\mathbf{x}v'},\ \beta\ (1 - r_{\mathbf{x}v'})\}$ where β is a positive coefficient and $r_{\mathbf{x}v'} = \mathrm{T}_1(\underline{\mathbf{x}},v')$ is the cross-correlogram between $Y_\mathbf{x}$ and $Y_{v'}$. Let r stand for its value at the origin (*change-of-support coefficient*): $r = r_{\mathbf{x}v}$ with $\underline{\mathbf{x}} \in v$. In such case, one has (Chilès and Delfiner, 1999, p. 411):

$$\forall p \in \mathrm{N}^*, \forall \underline{x}, v',\ \mathrm{T}_p(\underline{x}, v') = \frac{\Gamma(\beta)\Gamma(\beta r_{\underline{x}v'} + p)}{\Gamma(\beta r_{\underline{x}v'})\Gamma(\beta + p)} \tag{22}$$

The parameter inference and validation of the model are now detailed. For two different random samples $\{\underline{\mathbf{x}},\underline{\mathbf{x}}'\}$, Equation (21) gives:

$$\forall \underline{x} \in v, \underline{x}' \in v',\ \mathrm{T}_1(\underline{x},\underline{x}') = r\, r_{\underline{x}v'}\ ;\ \text{in particular, if } v = v',\ \mathrm{T}_1(\underline{x},\underline{x}') = \mathrm{T}_1^2(\underline{x}, v) = r^2\ .$$

Together with Equations (16) and (17), these formulae provide r and $r_{\mathbf{x}v'}$. The scalar parameter β can be chosen to honor Equation (4), whereas the correlograms of all the factors are obtained from Equations (14), (15) and (22). At this stage, the model is fully specified. It can be validated by comparing the variogram of the point-support Gaussian variable $Y_\mathbf{x}$ with its variograms of lower order, defined by

$$\forall \omega \in]0,2],\ \gamma_\omega(\underline{x},\underline{x}') = \frac{1}{2}\mathrm{E}\{|\, Y_{\underline{x}} - Y_{\underline{x}'}\,|^{\omega}\} \tag{23}$$

The usual variogram corresponds to $\omega = 2$, the madogram to $\omega = 1$ and the rodogram to $\omega = 1/2$. The Hermitian model satisfies the following relation (Emery, 2005):

$$\gamma_\omega(\underline{x},\underline{x}') = \mathrm{E}\{[1 - R(\underline{x},\underline{x}')]^{\omega/2}\}\frac{2^{\omega-1}}{\sqrt{\pi}}\Gamma\left(\frac{\omega+1}{2}\right) \tag{24}$$

In the stationary framework, this quantity only depends on the vector **h** separating the blocks containing **x** and **x**′. After simplification, the standardized variograms of order ω are expressed as a hypergeometric function (Slater, 1966) of the usual variogram:

$$\frac{\gamma_\omega(\mathbf{h})}{\gamma_\omega(\infty)} = {}_3\mathrm{F}_2(-\frac{\omega}{2}, \beta r, \frac{\beta}{r}[1 - \gamma_2(\mathbf{h})]; \beta, \beta; 1) \tag{25}$$

Equation (25) generalizes the relation between the madogram and the variogram given by Wackernagel (2003, p. 260). Note that it applies to the random-location variable $Y_{\mathbf{x}}$, hence it should be checked on the empirical variograms of $Y_{\mathbf{x}}$ for distances greater than the block size, for which the randomization of the samples within the blocks has almost no effect on the variograms.

5.3 OBSERVATIONS ON THE DISCRETE HERMITIAN MODEL

For mathematical consistency, it is preferable to define the change-of-support Hermitian model by reference to the point-block distribution, as in Equation (20). Indeed, let us write the general form of the Hermitian model for the point-point distribution (Eq. 5):

$$\forall p \in \mathsf{N}^*, \forall \underline{\mathbf{x}}, \underline{\mathbf{x}}', \ \mathrm{T}_p(\underline{\mathbf{x}}, \underline{\mathbf{x}}') = \mathrm{E}[R(\underline{\mathbf{x}}, \underline{\mathbf{x}}')^p] \tag{26}$$

where $R(\underline{\mathbf{x}},\underline{\mathbf{x}}')$ is a random variable in [-1,1]. Then, Equation (14) entails:

$$\forall p \in \mathsf{N}^*, \forall \underline{\mathbf{x}}, \underline{\mathbf{x}}' \in v, \ \mathrm{T}_p(\underline{\mathbf{x}}, v) = \sqrt{\mathrm{E}[R(\underline{\mathbf{x}}, \underline{\mathbf{x}}')^p]} \tag{27}$$

In general, these terms cannot be identified with the succession of moments of a random variable, hence the point-block *pdf* (Eq. 1) takes negative values and is inconsistent.

5.4 TWO LIMIT CASES: DISCRETE GAUSSIAN AND MOSAIC MODELS

These models correspond to the limit cases $\beta = \infty$ and $\beta = 0$ of the above beta model. In the discrete Gaussian model, $R(\underline{\mathbf{x}},v')$ is deterministic. The *change-of-support coefficient r* = $R(\underline{\mathbf{x}},v)$ with $\underline{\mathbf{x}} \in v$ can be determined i) from a variogram model of $Z_{\mathbf{x}}$ using Equation (4) or ii), from a variogram model of the Gaussian variable $Y_{\mathbf{x}}$ using Equation (17). In general, both alternatives are not fulfilled simultaneously, hence the discrete Gaussian model is over-determined and may not be internally consistent. Equation (17) would be helpful for the parameter inference in other isofactorial models, such as the gamma and Laguerre-type models (Hu, 1988; Chilès and Delfiner, 1999, p. 443).

Another limit case is the mosaic model, in which $R(\underline{\mathbf{x}},v')$ only takes two values: 0 or 1. In such case, Equation (20) shows that $\mathrm{T}_p(\underline{\mathbf{x}},v')$ does not depend on p. At a global level the block-support transformation ϕ_v is an affine function of the sample transformation ϕ (Eq. 3) and the change of support amounts to an affine correction. This is a limitation of the mosaic model, as the block-support distribution is expected to be less skewed than the point-support one. Concerning the local model, the correlograms $\mathrm{T}_1(\underline{\mathbf{x}},v')$ and $\mathrm{T}_1(v,v')$ obtained by Equations (14) to (16) are likely to be smooth near the origin, a feature incompatible with a mosaic distribution: $\mathrm{T}_1(\underline{\mathbf{x}},v')$ and $\mathrm{T}_1(v,v')$ must belong to the set of indicator correlograms, hence their behavior near the origin is at most linear (Matheron, 1989a, p. 22). In conclusion, more realistic models are needed to describe the change of support in the mosaic framework, both at the global and local scales. For instance, one can resort to mosaic-type models in which the cell valuations depend on their size (Matheron, 1989b, p. 317; Rivoirard, 1994, p. 27).

6 Case study

The previous concepts are illustrated on a real dataset from a Chilean porphyry copper deposit with 2,376 diamond-drillhole exploration samples measuring the copper grades. Each sample is a twelve-metre-long composite. The drillholes are located in an area of 400m × 600m × 130m (Figure 2A) and the copper grades are lognormally distributed with a mean value of 1.00% (Figure 2B). The structural analysis of the data reveals an anisotropy whose main axes are along the horizontal and vertical directions (Figures 2C and 2D). For ore reserve estimation, selective mining units of 15m × 15m × 12m are used. The purpose of the study is to model the grades at point and block supports using the discrete Hermitian model.

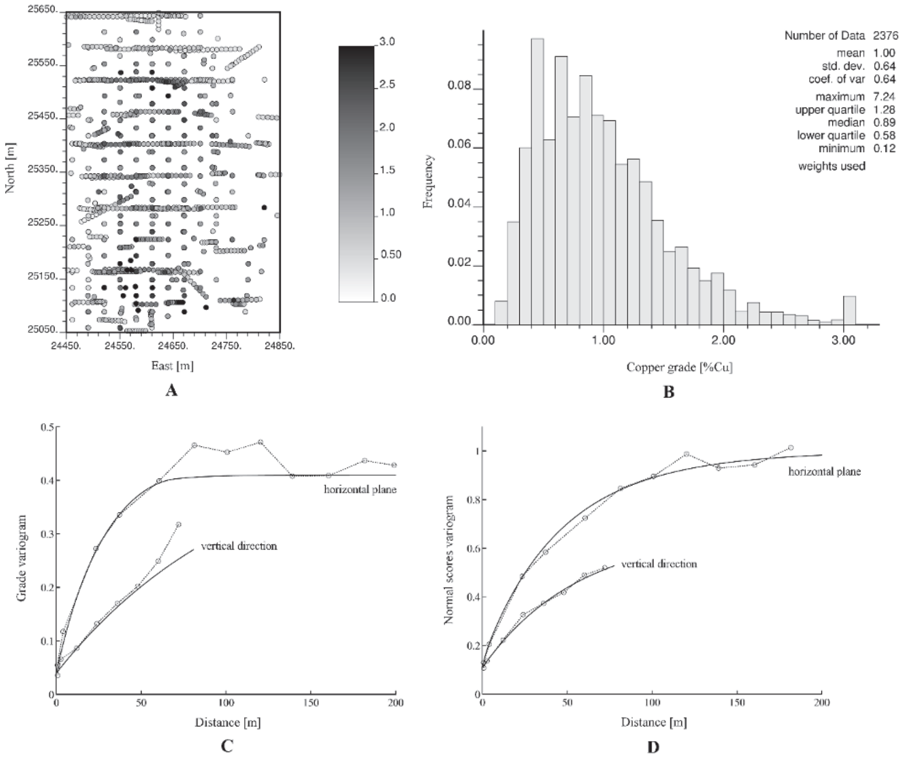


Figure 2. A, location map of the samples, B, declustered histogram, C, copper grade variogram and D, normal scores variogram

The variogram models for the raw variable (grade) and its normal scores transform are:

$$\gamma_Z(\mathbf{h}) = 0.04 + 0.13\,\mathrm{sph}(70\mathrm{m},190\mathrm{m}) + 0.24\exp(60\mathrm{m},240\mathrm{m})$$

$$\gamma_Y(\mathbf{h}) = 0.11 + 0.56\exp(170\mathrm{m},170\mathrm{m}) + 0.03\exp(5\mathrm{m},\infty) + 0.30\exp(100\mathrm{m},\infty)$$

A regularization of both variograms onto a 15m × 15m × 12m block support gives the change-of-support coefficient and the block-support variance:

$$\forall \underline{\mathbf{x}} \in v, \mathrm{T}_1(\underline{\mathbf{x}}, v) = 0.854 \text{ (Eq. 17) and } \operatorname{var}(Z_v) = 0.274 .$$

In the following, we assume that the point and block-support variables can be described by a Hermitian model with a beta random variable $R(\underline{\mathbf{x}},v')$ (section 5.2). The parameter β of this random variable is determined to honor the block-support variance (Eq. 4 and 22). The evolution of the block variance as a function of β, given $r = 0.854$, is shown in Figure 3A; for practical calculations, the expansion in Equation (4) is truncated at order $p_{\max} = 100$. The actual block-support variance (0.274) is obtained for a value of β close to 23. This value is validated by plotting the standardized experimental variograms of orders 0.5, 1 and 1.5 of the normal scores data as a function of their usual variogram, and comparing the experimental points to the theoretical curves of a beta model with parameter β = 23 (dashed lines in Figure 3B) (Eq. 25). In logarithmic coordinates, such curves are close to straight lines, hence a pure Gaussian model (corresponding to β = ∞) could also be used (Emery, 2005).

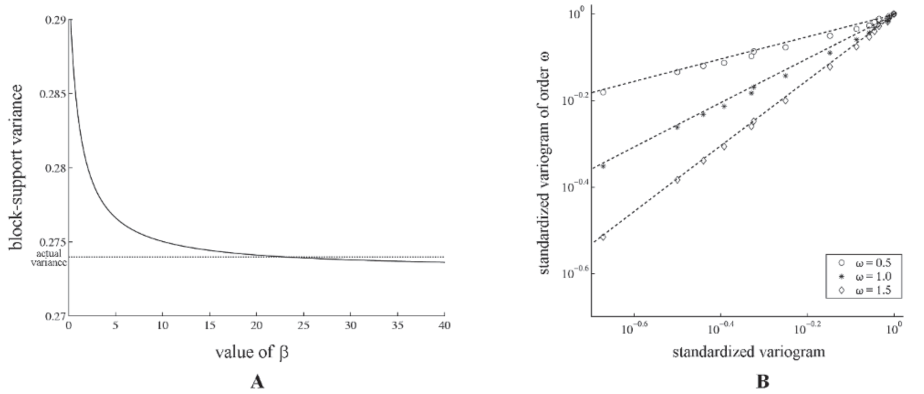


Figure 3. A, determination of parameter β and B, validation of the model

The analysis must be performed with care due to the sensitiveness of the parameter values to the histogram and variogram model:

- The point-support histogram and its variance are strongly dependent on the extreme values and the upper-tail modeling. If possible, one should fit a variogram with a sill that matches the histogram variance (i.e. the sum of the squared coefficients of the point-support transformation function, see Eq. 2), otherwise a shortcut solution consists in standardizing the variogram sill around this variance.

- The block-support variance depends on the variogram model at small distances (in particular the amplitude of the nugget effect), for which the data pairs are scarce. The same observation applies to the inference of the change-of-support coefficient from Equation (17).

- The coefficient β may be determined with low accuracy when its value is high (Fig. 3A); however, in this case, it has little influence on the results as the corresponding Hermitian model is close to the discrete Gaussian model.

7 Conclusions

This work focused on the inference and internal consistency of bivariate isofactorial models for change-of-support applications. A procedure has been proposed to improve the structural analysis and simplify it with respect to the traditional approach. In the Hermitian framework, the user should beware of both extreme cases (discrete Gaussian and mosaic models): the first one has no flexibility since a single parameter must fulfill two equations, while in the second one the global change of support amounts to an affine correction. The proposed beta model is more flexible and the parameter inference remains relatively simple. These results can be extended to other change-of-support models such as the Laguerre-type model.

Acknowledgements

The authors would like to acknowledge the Department of Mining Engineering at the University of Chile and its industry sponsor Codelco-Chile for supporting this research.

References

Chilès, J.P. and Delfiner, P., *Geostatistics: Modeling spatial uncertainty*, Wiley, 1999.

Emery, X., Conditional simulation of non-Gaussian random functions, *Mathematical Geology*, vol. 34, no. 1, 2002, p. 79-100.

Emery, X., Variograms of order ω: a tool to validate a bivariate distribution model, *Mathematical Geology*, vol. 37, no. 2, 2005, in press.

Emery, X. and Soto Torres, J.F., Models for support and information effects: a comparative study, *Mathematical Geology*, vol. 37, no. 1, 2005, p. 49-68.

Guibal, D. and Remacre, A.Z., Local estimation of the recoverable reserves: comparing various methods with the reality on a porphyry copper deposit, *in* G. Verly, M. David, A.G. Journel and A. Maréchal, eds., *Geostatistics for Natural Resources Characterization*, Reidel, vol. 1, 1984, p. 435-448.

Hu, L.Y., *Mise en œuvre du modèle gamma pour l'estimation de distributions spatiales*, Doctoral Thesis, Paris School of Mines, 1988.

Matheron, G., A simple substitute for conditional expectation: the disjunctive kriging, *in* M. Guarascio, M. David and C.J. Huijbregts, eds., *Advanced geostatistics in the mining industry*, Reidel, 1976a, p. 221-236.

Matheron, G., Forecasting block grade distributions: the transfer functions, *in* M. Guarascio, M. David and C.J. Huijbregts, eds., *Advanced geostatistics in the mining industry*, Reidel, 1976b, p. 237-251.

Matheron, G., The selectivity of the distributions and the "second principle of geostatistics", *in* G. Verly, M. David, A.G. Journel and A. Maréchal, eds., *Geostatistics for Natural Resources Characterization*, Reidel, vol. 1, 1984a, p. 421-433.

Matheron, G., Isofactorial models and change of support, *in* G. Verly, M. David, A.G. Journel and A. Maréchal, eds., *Geostatistics for Natural Resources Characterization*, Reidel, vol. 1, 1984b, p. 449-467.

Matheron, G., The internal consistency of models in geostatistics, *in* M. Armstrong, ed., *Geostatistics*, Kluwer Academic, vol. 1, 1989a, p. 21-38.

Matheron, G., Two classes of isofactorial models, *in* M. Armstrong, ed., *Geostatistics*, Kluwer Academic, vol. 1, 1989b, p. 309-322.

Rivoirard, J., *Introduction to disjunctive kriging and non-linear geostatistics*, Clarendon Press, 1994.

Slater, L.J., *Generalized hypergeometric functions*, Cambridge University Press, 1966.

Wackernagel, H., *Multivariate geostatistics: an introduction with applications*, 3rd Edition, Springer, 2003.

HISTORY MATCHING UNDER GEOLOGICAL CONTROL: APPLICATION TO A NORTH SEA RESERVOIR

B. TODD HOFFMAN and JEF CAERS
Department of Petroleum Engineering,
Stanford University, California, 94305-2220, USA

Abstract. Solutions to inverse problems are required in many Earth Science applications. The problem of determining reservoir properties, such as porosity and permeability from flow data, shortly termed "history matching", is one example. In many traditional inverse approaches, certain model assumptions are made on either the data likelihood or the prior geological model, e.g. assumptions of conditional independence between data or Gaussianity on the distributions, which do not reflect the reality of actual data. This limits the applications of such approaches to practical problems like history matching. While modeling assumption are inevitable, this paper presents a general inversion technique that can be used with different geostatistical algorithms to create models that honor several types of prior geological information and at the same time match almost any type the data. The technique is built on the idea of perturbing the probability distributions used to create the models rather than perturb the properties directly. By perturbing the probabilities, the prior geological model as described by a geostatistical model or algorithm is maintained. We present a practical implementation of the probability perturbation method. A case study demonstrates how the practical implementation would work in an actual situation. The case study is a North Sea hydrocarbon reservoir where the production rates and pressure information are iteratively included in the model.

1 Introduction

History matching is a term used in reservoir engineering to describe the problem of finding a 3D reservoir model that matches the observed production data. In that regard, the problem of history matching is no different from any other inverse modeling technique aimed at finding a set of model parameters, **m** based on measured data, **d**. The same issues need to be addressed:

- *Non-uniqueness*: many 3D models can be found that match equally well the production data.
- *Need for a prior geological model*: any of the models matching the production data should also honor information about the geological continuity, either provided by a variogram, object model or training image model. In inverse terminology: some information about the prior distribution of the model parameters, namely $f(\mathbf{m})$ is usually available.

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 1067–1076.

- *The forward model*: this model, further denoted as g: $\mathbf{d} = g(\mathbf{m})$, provides the relationship between the data and the model parameters (not considering data and model errors). g is often a strongly non-linear function, and in history matching it is provided by a flow simulator (finite difference/element model).

A problem specific to history matching problem lies in the model parameterization. Many reservoir features, such as fault position, fault transmissibilities, facies proportions, relative and absolute permeability, porosity etc… can be perturbed to achieve a history match. In most cases, large scale structures such as fault positions and layer geometries are adjusted by hand based on reservoir engineering expertise. The reservoir properties (facies, permeability and porosity) are modelled using geostatistical methods, hence need to be adjusted in an algorithmic fashion. In this paper we present a practical approach to the latter problem by means of the probability perturbation method. The theory behind this method will be briefly reviewed but is presented in greater details in other papers (Caers, 2003; 2004, this conference). This paper focuses on putting this method into actual practice by first extending the basic probability perturbation methods, then by presenting an actual reservoir case study.

2 Probability Perturbation Method

A brief explanation of the probability perturbation method (Caers, 2003; Caers, 2004, this conference) is given to provide the background for further development. For demonstration purposes, we will consider the case where the parameters **m** of the model are given by a set of binary spatial variables described by the indicator variables:

$$I(\mathbf{u}) = \begin{cases} 1 & \textit{if} \quad \text{the "event" occurs at } \mathbf{u} \\ 0 & \textit{else} \end{cases} \qquad (1)$$

where $\mathbf{u} = (x, y, z) \in$ model, is a spatial location, and $I(\mathbf{u})$ could denoted any spatially distributed event, for example, $i(\mathbf{u})=1$ means channel sand occurs at location **u**, while $i(\mathbf{u})=0$ indicates non-channel sand occurrence. An initial realization of $I(\mathbf{u})$ on the same grid containing all locations **u** will be termed $\mathbf{i}^{(0)}=\{i^{(0)}(\mathbf{u}_1) ,\ldots, i^{(0)}(\mathbf{u}_N)\}$. The method works equally well for continuous and discrete variables.

The prior model $f(\mathbf{m})$ in this paper is modelled using sequential simulation, whereby each variable $I(\mathbf{u})$ is simulated sequentially accounting for any linear data (hard data or soft data) and any previously simulated indicators. Each step in a sequential simulation algorithm consist of determining a local conditional probability distribution:

$$P(I(\mathbf{u})=1 \mid \text{previously simulated nodes + linear data})$$

or in shorter notation denoted as P(A|B). The initial realization does not match the non-linear production data **d**, hence need to be further perturbed in an iterative fashion.

Rather than perturbing the initial realization $\mathbf{i}^{(0)}$ directly, Caers (2003) proposes to perturb at each step of the sequential simulation algorithm, the probability model, P(A|B), used to generate the initial realization. This is done by introducing another probability model, P(A|D), where the notation D=**d** is used. The perturbation of P(A|B) by P(A|D) is achieved by combining both conditional probabilities using Journel's method (2002). The resulting new probability model, P(A|B,D), is used to draw a "perturbed" realization using the same sequential simulation algorithm, but with different random seed. P(A|D) is defined as follows:

$$\text{for all } \mathbf{u}\text{: } P(A|D) = (1-r_D)\, i^{(0)}(\mathbf{u}) + r_D\, P(A) \in [0,1] \qquad (2)$$

where r_D is a parameter between [0,1] that controls how much the model is perturbed. To better understand the relationship between r_D and P(A|D) consider the two limiting cases when r_D=1 and r_D=0. When r_D=0, P(A|D) = $i^{(0)}(\mathbf{u})$ and the initial realization, $i^{(0)}(\mathbf{u})$, is retained in its entirety, and when r_D=1, P(A|D) = P(A) and a new equiprobable realization, $i^{(1)}(\mathbf{u})$, is generated. The parameter r_D, therefore, defines a perturbation of an initial realization towards another equiprobable realization.

There may exist a value of r_D, such that $i^{(1)}{}_{\mathrm{rD}}(\mathbf{u})$ will match the data better than the initial realization. Finding the optimum realization, $i^{(1)}{}_{\mathrm{rD}}(\mathbf{u})$ is a problem parameterized by only one free parameter, r_D; therefore, finding the optimum realization is equivalent to finding the optimum r_D value.

$$r_{D_{opt}} = \min_{r_D}\{O(r_D) = \| D^S(r_D) - D \|\} \qquad (3)$$

where $O(r_D)$ is the objective function, which is defined as some measure of difference between the data from the forward model, $D^S(r_D)$ and the observed data, D. The value of r_{Dopt} and consequentially the optimum realization can be found using any one-dimensional optimization routine, for example the Brent method (Press et al., 1989).

3 Regional probability perturbation

The previously described method from Caers (2003) is theoretically well founded and shown to be linked to the well-established Bayesian inverse theory (Caers, 2004); however, going from theory to practice is a non-trivial step. A number of specific issues are addressed in the current paper, and they are summarized as follows:

(1) In many applications, the parameter space may be large. Hence, parameterizing the model perturbations using a single parameter may not achieve a satisfactory match in a reasonable amount of CPU time. Following, a higher order parameterization of the perturbation of **m** is proposed by dividing 3D space into regions, each with a different perturbation parameter r_D attached to it.

(2) A higher order parameterization leads to a more difficult optimization than the 1D optimization of Eq. (3). With discrete variables, such as in the binary case above, gradients may not be available, hence an efficient non-gradient approach is necessary.

3.1 Model Regions

The probability perturbation method (PPM) is able to perturb parameters and honor the conceptual geologic model; however, for models with vastly different properties in different parts of the model, the efficiency of the PPM is not satisfactory. For many applications, the models need to be able to account for local variability by perturbing parameters by different amounts in different regions in space (Hoffman and Caers, 2003).

To be able to achieve this, the region geometry must first be defined. The regions may be any arbitrary shape, but their definition is problem specific and will be left to the user. Some methods are discussed in the Case Study section. The regions are denoted as $\{R_1, R_2, \ldots, R_K\}$ where K is the total number of regions, and the entire realization is $R = (R_1 \cup R_2 \cup \ldots R_K)$. With multiple regions, P(A|D) must be defined slightly different than in the PPM. It continues to be defined for the entire reservoir, R, but its local value depends on the region definition: when **u** is located in region R_k, the perturbation parameter takes on a value of r_{Dk}. Therefore P(A|D) can have different values for different regions of the reservoir, and the following equation for P(A|D) is used.

$$P(A|D) = (1 - r_{Dk})\, i^{(0)}(\mathbf{u}) + r_{Dk}\, P(A) \qquad (4)$$

Each perturbation parameter is updated based on how well the model matches the data in each region. If the match is good, r_{Dk} is small or even zero, and if the match is poor, a value r_{Dk} close to one should be taken. Figure 2 shows an illustrative 2D fluid flow example of how this works.

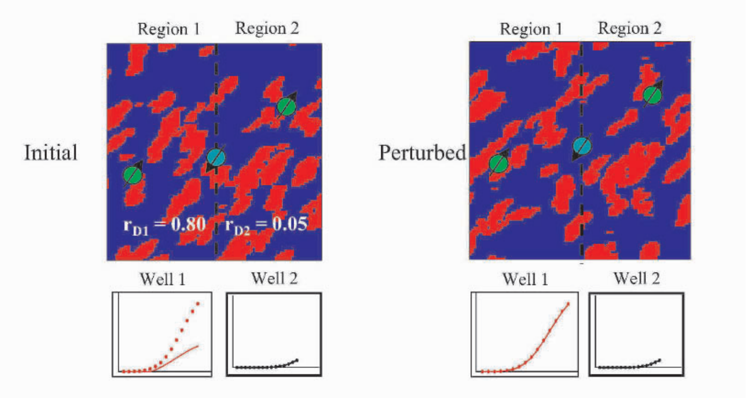


Figure 1: Perturbing two regions by separate amounts, without creating discontinuity at the border between the two regions.

There are three wells, one injector in the middle and two producers. In the initial model, the production well on the right (Region 2) is matching the production data (water cut in this case) quite well, whereas the well on the left (Region 1) is not matching nearly as well. Therefore, Region 2 will require a small perturbation parameter value, and Region 1 will require a larger perturbation. In the perturbed model, Region 2 is changed only slightly; the location of the bodies is roughly the same.

Conversely, the bodies location in Region 1 are considerably different in the perturbed model compared to the initial model.

Notice there are no model artifacts or discontinuities along the region border illustrating that the geology is always maintained. The reason for not creating artifact discontinuities can be explained by the nature of the sequential simulation algorithm and by the perturbation method applied. In sequential simulation, each grid block is simulated based on any reservoir data and on any previously simulated grid block properties. The method searches for any such previously simulated grid locations in an elliptical search neighborhood. This search neighborhood may (and should) cross the region-boundaries. When simulating a grid block in one region, the grid block properties in other regions are used to determine P(A|B,D), hence creating continuity across the boundaries. Secondly, geological continuity is assured in the perturbation method through the probability P(A|B), which is not calculated per region but for all regions together (Hoffman and Caers, 2003).

3.2 Optimization

Regional changes are convenient for local improvement in the data match; however, this requires multiple perturbation parameters, r_{Dks}, to be optimized. Because the forward model of these large problems may take up to several hours, a full multi-dimensional optimization is not feasible. Therefore, an efficient procedure to find the optimum values for r_{Dk} is developed. First, the forward model is completed on the entire model. Then the objective, O_k, is calculated for each region. The objective is simply the mismatch of the production data for all wells in that region.

$$\forall k = 1,\ldots,K \quad O_k = \left\| D_k^S - D_k \right\| \tag{5}$$

Because the data in each region can be influenced by the parameters outside that region, in general the values of O_k will depend on all $\{r_{D1} \ldots r_{DK}\}$. However, the data in a region is *principally* dependent on the model parameters within its region; thus, we will assume that O_k only depends on r_{Dk}. Based on this assumption, we can update the r_{Dks} in all regions based on the production data from one flow simulation. The next r_{Dk} for each region is determined by performing one step of a one-dimensional optimization routine, but since there are multiple regions, the "one step" must now be done K times (once for every region). After all r_{Dks} are updated, a new reservoir model is generated and flow simulation is completed. The optimization algorithm for the inner loop of the method is given:

1. Guess initial perturbation parameters (usually 0.5 for all regions).
2. Create new reservoir model and run forward model.
3. Calculate objective (mismatch) for each region.
4. For each region, perform one step of a 1D optimization to find the new perturbation parameters using Brent method (Press et al., 1989).
5. Goto step 2. Loop until there is no improvement in objective.

Note that the model is not broken into K independent problems where the forward model would be completed on each region separately. Rather, the realizations are always generated over the entire reservoir, and the forward model is always completed over the entire reservoir. Regions are only used for objective function calculations and perturbation parameter, $r_{Dk,}$ updating. The efficiency advantage lies in the fact that only one forward model is needed per iteration (the same as the single PPM), yet all regions can be improved during every iteration.

3.3 Perturbing prior proportions

The PPM and regional PPM perturb the parameters **m** and at the same time account for a prior geological model $f(\mathbf{m})$. An important part of the prior geological model is the marginal distribution on each parameter m_i. In the case of a binary variable as in Eq. (1), the prior model is the global proportion of the event occurring (e.g. the proportion of channel sand). In many practical situations, the global proportion is poorly known, since it needs to be estimated from limited data (wells). Moreover, this proportion may vary considerably over the extent of the reservoir and one may want to model a local proportion per region. Assuming a wrong local or global (prior) proportion may prevent the PPM or regional PPM from achieving a satisfactory match.

We propose to perturb the local proportions per region, LP_k, jointly with the parameters r_{Dk} using a coupled optimization as follows:

$$LP_k^{new} = LP_k^{old} + i_k r_D f_c, \quad k = 1,..., K \tag{7}$$

where k is the region indicator and K is the total number of regions. f_c is a user-defined constant that characterizes the amount of change allowed in each iteration. Since the values of r_{Dk} range from 0 to 1, when r_{Dk} equals 1, LP^{old} is either increased or decreased by an amount equal to f_c. The indicator term, i_k, determines whether the LP should increase or decrease and is defined as follows:

$$i_k = \begin{cases} 1 & \text{if increase in local proportion is desired} \\ -1 & \text{if decrease in local proportion is desired} \end{cases}$$

In many instances, there is a known relationship between the proportion of facies and the production (e.g. an increase in facies causes an increase in production), hence the value of i_k is known. However, in some instances the relationship is unknown (e.g. an increase in facies may cause an increase or a decrease in production), hence the direction of the perturbation (the value of i_k) is not known beforehand and needs to be calculated using a numerical gradient. Due to the nature of inverse problems involving discrete variables, only the sign of the gradient is used, and not the gradient magnitude.

4 Case Study

The case study is a prominent North Sea reservoir with 22 wells (14 producers and 8 injectors) and 5½ years of production data. There are four major horizons and the top horizon is isolated from the lower three horizons by an impermeable shale layer. A significant number of very low permeability nodules are found in the reservoir. They were created by the diagenesis of calcite and tend to have a lenticular shape. They typically have an areal extent of a few meters to tens of meters. Where clusters of these bodies are found, they can have a large affect on fluid flow in the reservoir.

4.1 Simulation (Forward) Model

The reservoir model is a structured stratigraphic model with 39 cells in the x-direction, 98 cells in the y-direction, and 41 cells in the z-direction, but only about half of those cells are active. There are just over 70,000 total active gridblocks in the model. The calcite bodies are relatively thin, so they are given no vertical thickness in the simulation model. Gridblock containing a body or a cluster of bodies will get a reduced or zero z-direction transmissibility. For this case, the regional proportion (*RP*) is not a volumetric value of proportion, but rather the proportion of gridblocks that have a reduced vertical permeability due to the presence of the calcite bodies. For example if the RP is 30%, this does not mean that 30% of a region's volume is calcite, instead, 30% of the gridblocks in the region have reduced vertical permeability.

The location and proportion of calcite bodies is uncertain, so they must be stochastically built into the reservoir model using the snesim algorithm (Strebelle, 2002). The training image used for the current work only needs to be 2D because these bodies are modeled without a significant vertical dimension. The size and shape of the bodies are not well known, but we assume that where clusters of bodies occur, they affect an area of minimum 0.04 km^2. On average the gridblocks have a length around 100 m in the x and y directions, so the size of the bodies in the training image is typically two gridblocks squared.

To perform history matching with the regional probability perturbation method, a method for defining regions in the reservoir is required. Streamlines are well suited for the job because they directly show the flow paths by which fluid enters a production well (Milliken et. al., 2001). These paths identify the gridblocks that, if changed, will have an obvious impact on a well's production. All blocks hit by the set of streamlines entering a well define the "drainage zone" for that well. The various drainage zones define the geometry of the regions used for history matching in this case study.

Water, oil and gas rates for the 14 production wells and RFT pressure data from a number of both injector and producer wells is available; however only water rate and RFT pressure are used in the objective function. In the simulation model, wells have fixed liquid rates; hence, if water rates are correct, oil rates are also correct as well as the water cuts. The rates are matched using monthly averages.

Parameters such as porosity, permeability, relative permeability and fault transmissibility were examined to determine if they should be perturbed in the history-matching algorithm. However, it was determined that the overriding factor in the model is the presence of calcite bodies. The calcite bodies are perturbed using the regional probability perturbation method. Both the locations and the regional proportions of the bodies are allowed to vary. The bodies are included in the simulation layers 11-36 of the 41 total layers.

4.2 Results

By perturbing only the calcite bodies and allowing all the other parameters to be the same as the initial model, a quality history match is achieved for both the rates and the pressures. The water rates for three wells are displayed in Figure 2, and the RFT pressure measurements for three different wells are shown in Figure 3. The black data (lines and dots) is the observed data, and the light gray data is the history matched results. The line with the crosses represents the water rate data from the initial model, and the open diamonds are the initial pressure data.

For well P-3, the initial model has a water rate and breakthrough time that is much too high and much too early compared to the observed data. The history matched breakthrough time is very close and the rate is improved significantly. For well P-4, water is breaking through too late in the initial model, but in the history matched model, the data is matching much better.

Wells P-3 and P-4 have the largest mismatches and thus show the greatest improvements, but other wells also improved. The initial matches were closer, but they still showed some improvement (e.g. Well P-10).

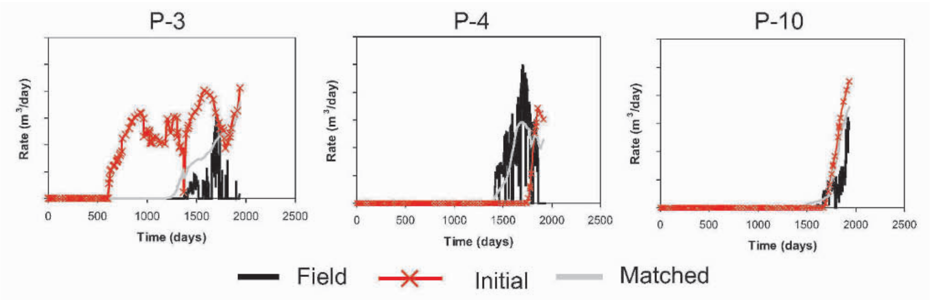


Figure 2: History match of water rates for three wells.

The pressure match was also improved. For some wells such as P-13, the pressure match from the initial model was already quite good, and that remained so in the history-matched model. Other wells such as I-6 went from a poor match to a very good match, and while well I-5 showed some improvement.

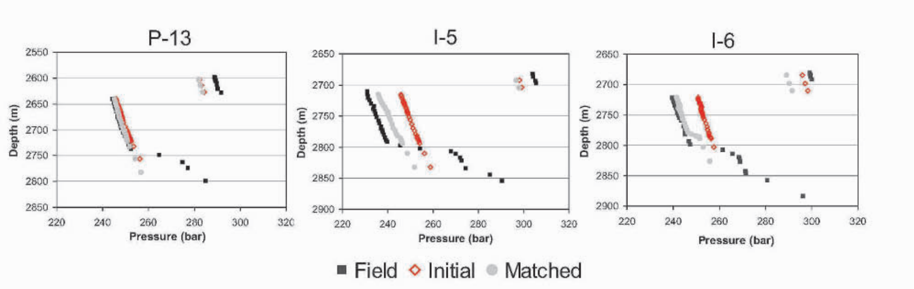


Figure 3: History match of pressures for three different wells.

For this North Sea reservoir example, the history matching procedure took 5 outer iterations. A normalized plot of the objective (mismatch) versus the iteration number is shown in Figure 4. Each iteration required 3 – 7 flow simulations, and the total number of flow simulation required is 28. The average run time for each simulation was about 2.5 hours, so the total CPU time was just under 3 days.

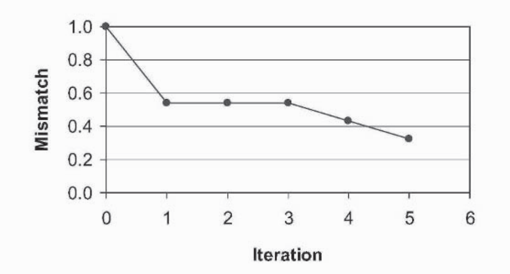


Figure 4: Normalized mismatch for history matching sequence.

Figure 5 shows the locations of the calcite bodies for two layers in a small segment of the reservoir. The proportion of bodies in the history matched model ranges from 1 % to 53 % with most regions having between 10 % and 20 %. The region with 53 % of its gridblocks affected by calcite bodies corresponds to well P-3. This well showed water breakthrough 2 years too early in the initial model, hence requiring a significant proportion of bodies to impede water flow.

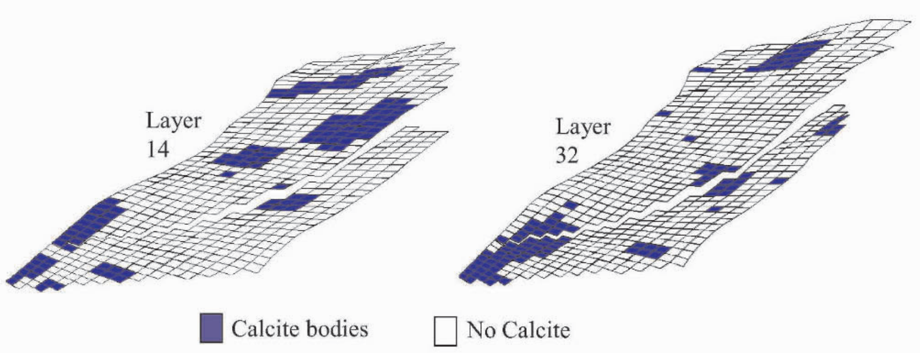


Figure 5: Location of calcite bodies in a history matched realization.

Conclusions

While the theory of the probability perturbation method has been established in other papers, this paper demonstrates the practicality for solving large complex inversion problems. History matching a North Sea reservoir serves as an example case study. The implementation is carried out by perturbing the locations and regional proportions of calcite bodies that have very low vertical transmissibility. More generally this work shows that large-scale structures such as the position of facies bodies as well as their local proportion can be perturbed in the inversion process.

Acknowledgements

We would like to thank Statoil for providing us with this dataset and the support from various engineers and geologists in the field and in the research center to help make the case successful. Specifically, we want to thank Martin Springer for his assistance preparing the data and answering our many questions. May Hege Viken, Alfhild Lien Eide and Magne Lygren also provided helpful insight into the work. We also want to acknowledge the Stanford Center for Reservoir Forecasting for their financial support of the research.

References

Caers, J., History matching under a training image-based geological model constraint. *SPE Journal*, SPE # 74716, 2003, p. 218-226

Deutsch, C.V. and Journel, A.G., GSLIB: Geostatistical Software Library and User's Guide, 2nd Edition, Oxford University Press, 1998.

Hegstad, B.K. and Omre, H. Uncertainty assessment in history matching and forecasting. In Proceeding of the Fifth International Geostatistics Congress, ed. E.Y Baafi and N.A. Schofield, Wollongong Australia, v.1, 1996, p. 585-596.

Hoffman, B.T. and Caers, J., Geostatistical history matching using the regional probability perturbation method. SPE Annual Conference and Technical Exhibition, Denver, Oct. 5-8. SPE # 84409, 2003, 16pp.

Milliken, W.J., Emanuel, A. S., and Chakravarty, A., Applications of 3D streamline simulation to assist history matching. SPE Journal RE&E, 2001, p. 502-508.

Journel, A.G. Combining knowledge from diverse data sources: an alternative to traditional data independence hypothesis. Math. Geol., v. 34, 2002, p. 573-596.

Press, W. H., Teukolsky, S.A., Vetterling, W.T., and Flannery, B.P.,Numerical Recipes in Fortran, Cambridge University Press, 1989, 963 pp.

Strebelle, S., (2002). Conditional simulation of complex geological structures using multiple-point geostatistics. Math. Geol., v34, p1-22.

A DIRECT SEQUENTIAL SIMULATION APPROACH TO STREAMLINE-BASED HISTORY MATCHING

J. CAERS, H. GROSS and A. R. KOVSCEK
Stanford University, Department of Petroleum Engineering, Stanford, CA 94305-2220

Abstract. Streamlines have proven useful for visualizing and solving complex history-matching problems that enhance reservoir description. This paper presents a novel, multiple-scale streamline-based technique that perturbs the reservoir model in a manner fully consistent with prior geological data. Streamlines define dynamic drainage zones around production wells and the mismatch between historical and simulated production is related to the average permeability of each zone. Direct sequential simulation is employed to propagate geological features within drainage zones. This method does not rely on time-of-flight inversion, nor is a tedious multidimensional optimization problem solved.

1 Introduction

History matching plays an important role in monitoring the progress of oil-recovery displacement processes, predicting future recovery, and choosing possible locations for the drilling of infill wells. While it is possible to formulate a history-matching algorithm in a general form consistent with inverse theory and constrained to a prior model, the shear number of unknowns to be estimated, combined with the complexity of the forward process model, make such an effort daunting. For instance, a small reservoir model might contain 50,000 to 100,000 grid blocks; in every grid block the permeability may be a model parameter. Physically-grounded inversion techniques help to resolve nonuniqueness and aid in the formation of algorithms that complete in acceptable time. Various data sources such as geological and seismic interpretations are available as constraints to the inverse flow problem, e.g, (Landa and Horne, 1997; Wang and Kovscek, 2003). In general, most approaches have been formulated to honor the histogram generated from sparse measurements of permeability and are restricted to variogram-based geological models (Caers et al., 2002).

This paper illustrates the methodology and basis for an extension to our previous streamline-based history matching efforts under geological constraint (Wang and Kovscek, 2000; Caers et al., 2002). Elsewhere, application of this technique to a large, complex, mature reservoir is reported (Gross et al., 2004). Our primary

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 1077–1086.

tools are incorporation of streamline information into the inversion methodology and a prior model formulation using direct sequential simulation–DSSIM (Journel, 1993). The permeability variogram available from measurements is honored from iteration to iteration. DSSIM does not require the explicit specification of a permeability histogram, a property that is used to our advantage. Core permeability measurements are generally taken on a scale (core scale) unrepresentative of the modeling scale. Constraining the prior model to a fixed permeability histogram therefore unduly restricts the model perturbations to a possibly wrong marginal distribution. Instead, the combination of streamline inversion and DSSIM with a locally varying mean–LVM is explored as a new approach to history matching.

2 Streamline-Based Inversion

A streamline is tangent everywhere to the instantaneous fluid velocity. Streamlines bound streamtubes that carry fixed volumetric flux when the system is incompressible. In this approach, flow rate is assigned to streamlines (Batycky et al., 1997). The time of flight, τ is the time required for a volume of fluid to move from the start (injector) to the end (producer) of a streamline. In a sense, τ indicates the breakthrough time for a streamline and the water-cut (water produced upon total fluid produced) versus time curve for a producer represents the sum of the production of all streamlines (Wang and Kovscek, 2003). Streamline simulation (in a forward sense) assumes that displacement along any streamline is one-dimensional and that streamlines do not interact. Thus, the flow problem is decomposed into a series of one-dimensional flow simulations linked by common boundary (i.e., well) conditions (King and Datta Gupta, 1998).

2.1 BASICS OF STREAMLINE INVERSION

With respect to history matching, casting the inversion problem within a streamline framework has many advantages. Each streamline carries a small portion of the injected fluid and the breakthrough of injected fluid at production wells is associated to individual streamlines (Wang and Kovscek, 2000). Moreover, non-interacting streamlines and incompressible fluids match the assumptions of the Dykstra-Parsons (1950) method for heterogeneous, layered porous media. Derivatives of the breakthrough time of any streamline, or subset of streamlines, are thereby approximated analytically and efficiently.

Any existing streamline inversion method, perhaps, can be used; however, we use exclusively the methodology of Wang and Kovscek (2000). The approach, first, relates the error

$$E_t = \frac{1}{N_{SL}} \sum_{m=1}^{N_{SL}} E^2_{t,BT,m} \tag{1}$$

$$E_{t,BT,m} = t_{D,BT,m} - t^0_{D,BT,m} \tag{2}$$

between measured,$t^0_{D,BT,m}$, and computed, $t_{D,BT,m}$, water breakthrough of each streamline to the effective permeability, k_{SL_m}, along each streamline. Note that

$t_{D,BT,m}$ signifies the breakthrough time of streamline m and N_{SL} is the number of streamlines. The term ”effective permeability” is used to signify a flow-weighted average permeability along a streamline:

(3)

where $u_j = (x_j, y_j, z_j)$ are the coordinates of streamline j, τ_m is the total time of flight, τ_{mj} is the time of flight through the grid cell u_j, and the summation is taken over the permeability, $k(u_j)$, of each grid block through which streamline m passes. The second step, discussed shortly, propagates the permeability perturbation to the underlying grid.

In the limits of unit mobility ratio, incompressibility, and a large number of streamlines, N_{SL}, Wang and Kovscek (2000) reduce the minimization problem to

$$J^T \Delta k_{SL}^R = -E \tag{4}$$

where the Jacobian, J, is diagonally dominant

$$\begin{array}{rcll} J_{mj} = & -1 & m = j & (5) \\ & \frac{1}{N_{SL}} & m \neq j & (6) \end{array}$$

and E is a vector containing the error for each streamline, Eq. 2. The vector, Δk_{SL}^R contains the relative perturbations to streamline effective permeability to attain a match:

$$\Delta k_{SL}^R = \frac{\Delta k_{SLm}}{k_{SLm}} \tag{7}$$

2.2 STREAMLINE-DERIVED FLOW ZONES

Streamline simulation is often appreciated for its fast and efficient computational properties. Streamlines also allow visualization of flow through the reservoir. With streamline trajectory one identifies: (i) the fraction of the entire field contacted with injected fluid, (ii) the volume of the reservoir drained by a given producer, (iii) the reservoir volume affected by a particular injector-producer pair, and (iv) the trajectory and volume associated with a given streamline. The definition of flow zones *a priori* is difficult and needs to be determined during every iteration of the history-matching process. To define the producer zone of influence, we track the set of permeability values through which each streamline passes (Emanuel and Milliken, 1998), as illustrated by Caers (2003). All streamlines entering a particular producer are grouped together, thereby defining a portion of the reservoir. There are as many producer flow zones as there are producers.

An advantage of locating the volume drained by a producer is that an estimate of the average change to the permeability of that flow zone is given by Eq. 7, once all streamlines in a flow zone are grouped together. By choosing to correct using information from individual streamlines, or groups of streamlines, the correction scale is modulated. Details follow in the presentation of the algorithm.

3 Prior Model Formulation Using DSSIM with LVM

Sequential Gaussian simulation (Deutsch and Journel, 1998) is the most widely used sequential simulation algorithm. All measured data are transformed into a standard Gaussian space. The entire simulation then takes place in standard normal space, and back transformations are finally performed using the histogram of the original permeability data. Normalization of simulated data to the original histogram is valid only when the histogram is known accurately *a priori*. For history-matching purposes, the histogram is not known with sufficient accuracy and constraining an inversion to such data results in an overly narrow search space for permeability.

Direct sequential simulation (Journel, 1993) appears to be better suited for history matching purposes. DSSIM is a particular form of sequential simulation where no transformation into Gaussian space is required. The simulation takes place directly in the data space, and an explicit histogram (marginal distribution) need not be provided to DSSIM. Inversions are constrained to the permeability variogram ensuring geologically sound solutions to the inverse problem.

The theoretical foundations of DSSIM are given by the sufficient conditions of honoring kriging means and variances at each node to be simulated along the sequential simulation path. This ensures reproduction of the variogram. For each node, the permeability value is drawn directly from a local conditional distribution type specified by the user. Any type of distribution is allowed, and it need not be stationary over the field. The advantage of using DSSIM lies in the fact that the histogram is not fixed *a priori*. The ability to allow the histogram to be perturbed, while at the same time honoring the variogram, provides a great deal of flexibility, for instance through perturbation of a local mean, to improve the match accuracy.

We employ DSSIM with locally varying means to perturb portions of the reservoir to achieve a better match, without creating discontinuities on the edges of the perturbed zones. In a hierarchical sense, local means are used to perform corrections to permeability at the field scale through alterations of the overall field mean. At the producer scale, corrections are performed by altering the mean permeability of the producer flow zone. LVM is used to perform corrections at integral scales, and thus transmit local corrections computed with streamlines to final permeability fields (Caers, 2000). Because the local mean is now provided by the correction arising from the streamlines, Eq. 7, the permeability histogram changes during every iteration.

4 Proposed Algorithm

The goal of the algorithm is obtain a match to production history by modifying the permeability field:

1. Choose an initial permeability field, $k(\mathbf{u})$ that is consistent with the permeability histogram and variogram and has an initial global mean and variance.
2. Iterate from $l = 1$ to $l = L_{max}$ or until convergence

a. Run a flow simulation to obtain the production history of all producers and a map of streamline trajectories. Streamline maps are updated as frequently as any major change in well condition. Alternately, the pressure of every grid cell may be output and the streamline trajectory calculated directly (Pollack, 1988).
b. Identify the flow zone (i.e., reservoir volume) associated with a given producer (Caers, 2003) for all streamline maps.
c. Average the flow zone data using all streamline maps to obtain a time-averaged flow zone.
d. Calculate the mismatch between simulated and measured history for every well, j, as

$$\Delta Q_{o,j} = \sum_{i=2}^{N_{ts}} \frac{\left|Q_o^{sim}(t_i) - Q_o^{hist}(t_i)\right| + \left|Q_o^{sim}(t_{i-1}) - Q_o^{hist}(t_{i-1})\right|}{\left|Q_o^{hist}(t_i) - Q_o^{hist}(t_{i-1})\right|} \tag{8}$$

where Q_o is the oil production rate, the superscripts *sim* and *hist* refer to simulated and measured data, respectively, and N_{ts} refers to the number of time steps. To date, the mismatch is gauged only on oil production rate.
e. Calculate the change in the average permeability, Δk_j^R for each producer flow zone as

$$\Delta k_j^R = \frac{1}{N_{ts}} \sum_{i=1}^{N_{ts}} \left(\frac{Q_{w,i}^{sim}(t)}{Q_T^{sim}(t)} - \frac{Q_{w,i}^{hist}(t)}{Q_T^{hist}(t)} \right) \tag{9}$$

This is equivalent to lumping all streamlines entering a producer into a single streamline and then applying the method of Wang and Kovscek (2000).
f. The change in permeability, Δk_j^R is taken as the mean change in permeability required for each region and this change is propagated to the grid using DSSIM. Return to step 2a.

5 Case Study

Each step of the history-matching technique is described through a validation exercise on a synthetic field with a limited amount of production data. The synthetic reservoir is 5000x5000x1000 ft discretized on a 100x100x10 grid. The large thickness of the reservoir implies that gravity is a driving force in the recovery process. A view of the reference permeability field and the relative permeability curves is given in Fig. 1. The permeability distribution of the field is close to lognormal with an average around 1000mD, the standard deviation is close to 300 mD, and has values ranging from 5mD to 2500 mD. The oil-water relative permeability curves are not adjusted during matching. The average porosity is 0.22 with a small variance (min. 0.19, max. 0.24). Porosity and permeability values are assumed uncorrelated.

The field has 10 injectors and 10 producers (Fig. 2); all are on during the 10 years of production. The field is initially at 5,000 psi, and there is no mobile water.

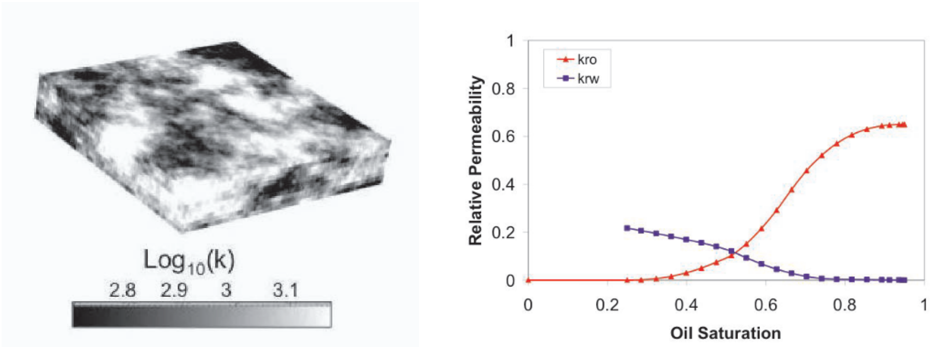


Figure 1. Reference data: permeability field (left) and oil-water relative permeability functions (right). The synthetic reservoir is 5000x5000x1000 ft.

Injection rate is set to 1,000 STB/d for all injectors and producers are set on BHP constraint of 4600 psi. A small solution gas-oil ratio simulates a reservoir containing dead oil. The end-point mobility ratio is unfavorable ($M = 8$). Production data were collected for oil (STB/d) and water (STB/d), and rates range between 100 and 1,000 STB/d per well. The field originally has 77.6 MMSTB in place, and after 10 years of production, 56.9 MMSTB are left; thus, 27 % of the oil in place has been produced. The streamline simulator 3DSL (Batycky et al., 1997) is run in incompressible mode for inverse calculations requiring 8 min of CPU time on a 1.8 GHz PC. In total, 8,500 streamlines and 8,500 gravity lines are updated every 3 months during the simulation. Streamline maps are displayed every 2 years. A typical streamline map is shown on Fig 2. From the identification of the streamlines entering a producer, the field is divided into producer flow zones.

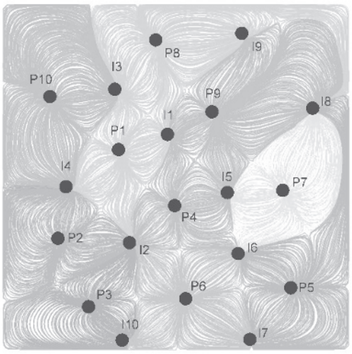


Figure 2. Areal view of positions of 10 vertical injectors (I) and producers (P). Streamlines indicate the division of the field into producer-based flow zones.

5.1 MATCH

Once an initial permeability field is set up and reference production history obtained, the initial permeability field is modified automatically to retrieve reference production curves and an estimation of the true permeability field. The permeability variogram is used to constrain updates to the permeability field.

Although a match between historical and simulated rates is the primary target of history-matching, restoring the reference permeability field is a more powerful indicator of the predictive power of the technique. In fact, being able to retrieve an accurate estimate of the reference permeability field starting from an initial permeability field and historical production rates indicates that the technique provides permeability model enhancements. Figure 3 illustrates the initial permeability field input to the history matching routine as well as the final distribution of permeability. Eight iterations were required to obtain convergence. Direct visual comparison of the final field in Fig. 3 with the reference field in Fig. 1 is difficult. Accordingly, a difference map between these permeability fields was prepared, Fig. 4. The map is obtained by subtracting the history-matched field from the reference field on a grid-cell by grid-cell basis. Figure 4 also presents a summary of the relative error between the reference production data and the production obtained with the final permeability model. Interestingly, initial sharp differences in contrast maps tend to smooth out gradually, meaning that the technique employed here converges towards the reference permeability field. Seven out of ten producers show significant reduction in error.

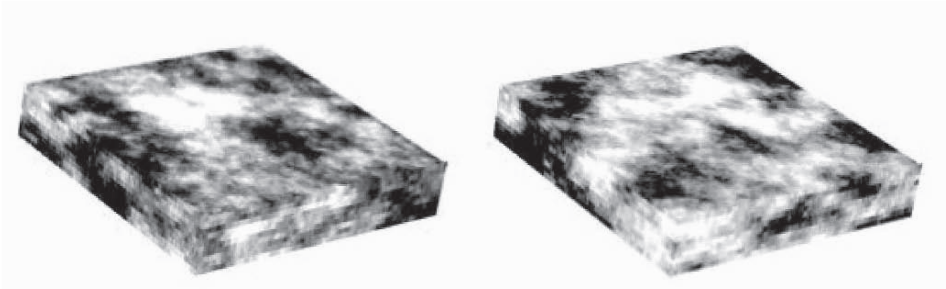


Figure 3. Initial permeability field for history matching (left) and final permeability field after history matching efforts (right). Gray-scale shading is identical to Fig. 1.

Only one final permeability field is presented, for reasons of brevity. Few iterations and relatively short time are required to obtain a satisfactory match. Although not proven, we assert that the technique allows of order ten to hundreds of matched models to be obtained rapidly. Multiple history-matched models are obtained (by changing random seeds in the sequential simulation) thereby allowing a statistical treatment of the matched model and predictions of future production. Finally, Fig. 5 compares the variograms for reference, initial, and final reservoir models. The variogram from the history-matched case is identical to the reference. Geostatistical consistency is maintained at all levels.

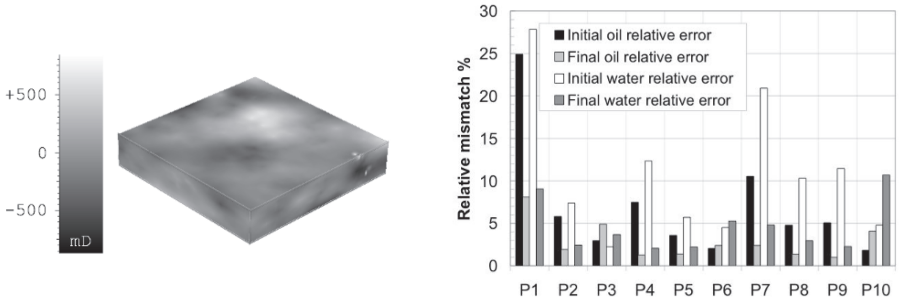


Figure 4. Summary of history matching efforts: difference map (left) between initial and reference permeability fields and improvement in match to historical production data (right).

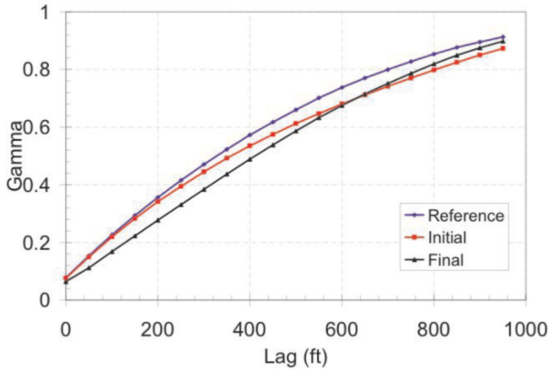


Figure 5. Semivariogram of reservoir models.

6 A Multi-Scale Perturbation Method

In many practical cases DSSIM with LVM provides a reasonable match to production data (Gross et al., 2004). The algorithm presented above perturbs the full permeability field by perturbing the locally varying mean. The random seed that is used to generate the single permeability field is fixed. In this approach, it is assumed that production data informs the local mean variations in permeability and is not sensitive to any small scale variation of the permeability. If in addition to a local mean variation, small scale variations in permeability have an impact on the reservoir production, then perturbing the local mean only may not achieve a history match. In addition to the local mean, the small scale permeability variation within each streamline region needs to be perturbed. We propose a multi-scale approach that perturbs both the local mean (coarse scale) and the underlying fine-scale permeability. The fine-scale permeability is perturbed using the (PPM) probability perturbation method (Caers, 2003). In short, PPM achieves a perturbation between an initial guess realization and another equiprobable realization and the perturbation is characterized by a single parameter. PPM can be applied to any sequential simulation approach including DSSIM with LVM. Figure 6 shows that the initial guess realization and any of the perturbations are constrained to a locally varying mean and a given permeability variogram. The magnitude of the parameter r_D defines the amount of perturbation. For any locally varying mean derived from streamlines, the parameter r_D is found by solving the following optimization problem:

$$O(r_D) = \sum_m \sum_t \left(k^{\circ}_{SL_m}(t) - k_{SL_m}(u, r_D) \right)^2 \tag{10}$$

where $k^{\circ}_{SL_m}$ is the desired streamline (coarse scale) effective permeability resulting from application of Eq. 7.

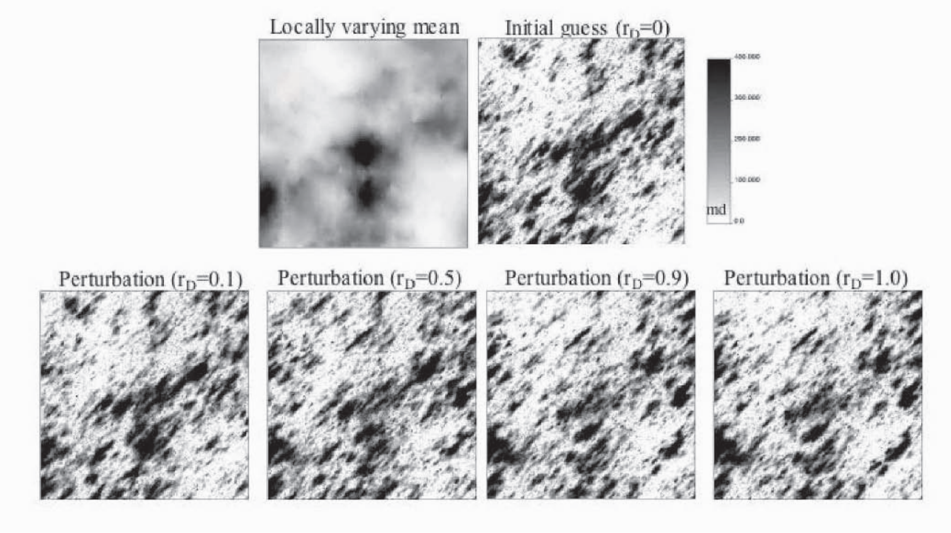


Figure 6. Example of multiscale correction based on LVM and employing coarse and fine scale perturbations. Shading represents permeability in md.

7 Summary

A multiscale history-matching algorithm is developed featuring large to small scale corrections. The technique is well suited to problems with numerous injectors and producers. Geostatistical data are honored at all stages of the algorithm, and, thus, the matches obtained remain consistent with the permeability variogram and hard data. The advantages of the technique include: (i) adaptability–it is applicable to any geological situation described by geostatistics and modeled by streamline simulation, (ii) speed–usually less than 10 iterations are needed to reach a satisfactory match, (iii) predictive power–all realizations honor the prior geological model. The drawbacks are similar to other history-matching techniques: (i) convergence–as in all gradient based techniques, the convergence depends on the initial guess, (ii) complexity–streamline and geostatistical concepts must be employed properly, (iii) prior information–predictive power relies on accurate measurement of production and fluid properties as well as a robust and accurate prior geological model.

Acknowledgements

We thank the SUPRI-A and SCRF Industrial Affiliates for financial support. Additionally, we acknowledge StreamSim Technologies Inc. for providing access to 3DSL.

References

Batycky, R. P., Blunt, M. J., and Thiele, M. R. A 3D Field-Scale Streamline-Based Simulator, *Soc. Pet. Eng. Res. Eng.*, 13(4) Nov. 1997, 246-254.

Caers, J., Adding Local Accuracy to Direct Sequential Simulation, *Math. Geol.*, 32(7)Oct. 2000, 815-850.

Caers, J., Krishnan, S., Wang, Y., and Kovscek, A. R., A Geostatistical Approach to Streamline-Based History Matching, *Soc. Pet. Eng. J.*, 7(3) Sept 2002, 250-266.

Caers, J., Efficient Gradual Deformation Using a Streamline-Based Proxy Method, *J. Pet. Sci. & Eng.*, 39 2003, 57-83.

Caers, J., History Matching Under a Training Image Based Geological Model Constraint, *Soc. Pet. Eng. J.*, 8(3) Sept 2003b, 218-226.

Deutsch, C.V. and Journel, A.G., *GSLIB: Geostatistical Software Library and User's Guide, 2nd Edition*, Oxford University Press, 1998.

Dykstra, H. and Parsons, R.L. The Prediction of Oil Recovery by Waterflood, Secondary Recovery of Oil in the United States, Principles and Practice, American Petroleum Institute, Dallas, TX, 1950, 160-174.

Emanuel, A.S. and Milliken, W.J., History Matching Finite Difference Models With 3D Streamlines, SPE 49000, Proceedings of the Annual Technical Conference of the SPE, New Orleans, LA, 27-30 Sept. 1998.

Gross, H., Thiele, M.R., Alexa, M.J., Caers, J.K., and Kovscek, A.R., Streamline-Based History Matching Using Geostatistical Constraints: Application to a Giant, Mature Carbonate Reservoir, SPE 90069, Proceedings of the Annual Technical Conference of the SPE, Houston, TX, 26-29 Sept. 2004.

Journel, A.G. Geostatistics: Roadblocks and Challenges, In Soares, A (ed.) *Geostatistic–Troia*, Kluwer, Dordrecht, 1993, 213-244.

King, M. J. and Datta Gupta, A. Streamline Simulation: A Current Perspective, *In Situ*, 22(1) 1998, 91-140.

Landa J.L and Horne, R. N. A Procedure to Integrate Well Test Data, Reservoir Performance History, and 4D Seismic Data Into A Reservoir Description, SPE 38653, Proceedings of the Annual Technical Conference of the SPE, San Antonio, TX, 5-8 Oct. 1997.

Pollack, D.W. Semianalytical Computation of Path Lines for Finite Difference Models, *Ground Water*, 26(6) 1988, 743-750.

Wang, Y. and Kovscek, A. R., A Streamline Approach for History Matching Production Data, *Soc. Pet. Eng. J.*, 5(4) Dec 2000, 353-362.

Wang, Y. and Kovscek, A. R., Integrating Production History into Reservoir Models Using Streamline-Based Time-of-Flight Ranking, *Petroleum Geoscience.*, 9 2003, 163-174.

MAPPING ANNUAL NITROGEN DIOXIDE CONCENTRATIONS IN URBAN AREAS

DAVID GALLOIS[(A)], CHANTAL DE FOUQUET[(A)], GAELLE LE LOC'H[(A)],
LAURE MALHERBE[(B)], GIOVANNI CARDENAS[(B)]
(A)Ecole des Mines de Paris - Centre de Géostatistique
35, rue saint-Honoré - 77305 Fontainebleau, France.
(B) INERIS. Parc technologique ALATA-60550 Verneuil-en-Halatte, France

Abstract

Urban background nitrogen dioxide (NO_2) is measured using passive samples, exposed during several consecutive fortnights in winter and in summer. Because of unavoidable technical incidents, the total number of "annual" measurements collected is limited to a few tens, which is not sufficient for estimating precise maps.
NO_2 comes mainly from the combustion of fossil hydrocarbons. Auxiliary variables like emission inventories, population density or land use, giving an approximate description of those emitters, may be entered in the mapping process as additional information.
For two French cities (Mulhouse and Montpellier) with different geographic context, the relationships between seasonal NO_2 concentrations and auxiliary variables are thoroughly examined. A high correlation between seasonal concentrations is shown, as the difference of spatial structures consistency for winter and summer concentrations.
The usual cross validation method brings out the interest of cokriging the annual concentration from the seasonal measurements, with auxiliary variables as external drift. This approach ensures the consistency of seasonal or yearly concentration estimations and allows a greater precision by the use of all available measurements.

1 Introduction

Nitrogen dioxide NO_2 is an urban air pollutant formed by reaction of oxygen and nitrogen produced by the combustion of fossil hydrocarbon. The main sources are road traffic, heating and specific industrial activities. Due to complex meteorological and photochemical phenomena, NO_2 increases in winter and is lower in summer. As NO_2 can cause respiratory irritations, the European regulation fixed an annual mean lower than 40 $\mu g/m^3$ as quality objective for 2010.
Nowadays, permanent stations measuring air pollution are only few per town, and it is not possible to obtain a precise cartography of the yearly or seasonal means from those measurements. Thus, monitoring campaigns are conducted in some towns, first to characterize the concentrations level in relation to the main pollution sources (main road, industrial zones, etc.), second to map NO_2 yearly mean levels as precisely as possible. During these campaigns, NO_2 is measured using "passive diffusion samples",

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 1087–1096.

installed at sites carefully chosen as representative of the background pollution, and exposed for several successive fortnights in winter and in summer.

NO_2 measurement campaigns being expensive, the objective assigned to geostatistical studies is to improve the accuracy of the estimation by taking into account additional information, providing an approximate description of the emission sources. For example, road traffic and heating should be partially linked to the population density or the land use, e.g. residential, industrial… and the local density of building. For some agglomerations, emission inventories including an evaluation of the local road traffic, or the declaration of industrial emissions made by the firms, are available.

When this auxiliary information is known at local scale, for example over a 1-km resolution grid, it can be used as external drift, or in a cokriging process. Bobbia et al. (2000) presented an instructive comparison of concentration maps estimated with or without auxiliary information. The remaining question was then to choose for each case the "best auxiliary variables", sometimes among a lot of information.

As measurement campaigns last several fortnights, it occurs that some of the "seasonal" measurements are missing because of technical problems. In this case, it is well known (Matheron, 1970) that cokriging the yearly concentration from the seasonal measurements allows using all the available seasonal data and ensures the consistency of the estimations, provided the multivariate variogram model between seasonal and yearly concentrations is consistent. Cokriging the yearly concentration is then equivalent to cokriging each seasonal concentration and calculating their average. In addition the first one gives the cokriging variance of the yearly mean. When the concentrations are separately estimated by kriging, the consistency between seasonal and yearly estimations is no more ensured, except in some very particular cases. The interest of cokriging will be shown on an example.

Despite not to be neglected (Gallois, 2004) the time component of the estimation variance is not considered in the present paper, focusing only on the spatial estimation.

2 Brief literature review

The European Framework Directive on Ambient Air Quality Assessment defines a regulatory framework for monitoring and evaluating air quality. Air pollution mapping at a relevant temporal scale is a valuable tool for providing the required information. In that context, geostatistical methods have been receiving particular attention for a few years and are now commonly applied by the French air quality monitoring (AASQA).

Kriging techniques, which were rather used to interpolate concentrations in areas equipped with a relatively dense monitoring network (Casado & al., 1994; Lefohn & al., 1988;Liu et al., 1996;Nikiforov et al., 1998; Tayanç, 2000), have been implemented to process data from passive sampling campaigns at an urban or regional scale. Pollutants under study are especially ozone, nitrogen dioxide and benzene. From a spatial point of view, interest has been paid to multivariate estimation, such as external drift or cokriging, to introduce auxiliary information in the estimation process (Bobbia et al., 2001;Cardenas et al., 2002; Phillips and al., 1997). Such methods may significantly improve the results provided that the auxiliary variables are properly chosen.

3 Correlation of NO_2 concentrations : seasonal values and auxiliary information

In the following, neglecting the time component of the estimation error, we assimilate the average of the three or four fortnightly measurements collected a season with the seasonal concentration, and their average with the yearly concentration.

The exploratory data analysis shows that for two agglomerations with very different geographical and industrial context, NO_2 concentrations present several analogies, mainly the expected links with some auxiliary variables. A useful and new result is the high correlation between seasonal concentrations.

3.1 MEASUREMENTS DURING THREE SEASONS IN MONTPELLIER

3.1.1 Urban context and sampling

With about 400 000 inhabitants, the agglomeration of Montpellier is located in the south of France. Close to the Mediterranean Sea, it is exposed to strong winds, and presents some relief. Three sampling campaigns of four fortnights each were conducted in different parts of the agglomeration in winter 2001, summer 2001 and winter 2002.

In addition, several auxiliary variables are available:

- Population density, averaged on discs centred on the sample location, with a radius of 200m, 1000m and 1500m;
- Nitrogen oxides emission inventory NO_X, including only road traffic evaluations, given on a grid with kilometric mesh.

The location of the samples varies a lot from one season to another. Among 143 samples site, 25 are common to winter and summer 2001, and 21 common to winter 2002 and summer 2001; only 3 are common to both winters among which 2 are common to the three seasons (Figure 1.).

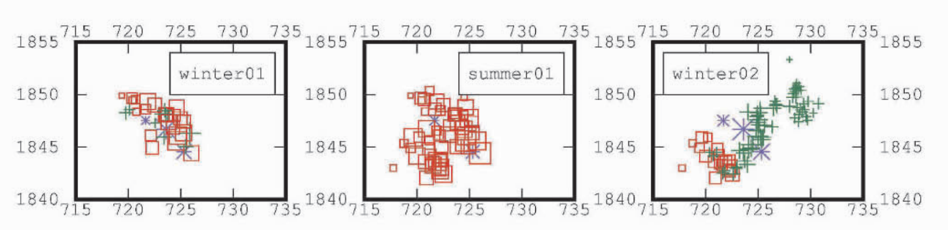


Figure 1. Location maps of NO_2 concentrations in Montpellier. *: informed in both winters; □: informed in summer; +: others. The presented area is identical for the three maps. Distances are given in km. The symbol size is proportional to NO_2 concentration.

3.1.2 Relationship between NO_2 concentration and auxiliary information

The three sampled zones correspond to different urban environments (Figure 2): the mean of population density on sampled sites decreases from about 5600 ha/km^2 for winter 2001 in the urban centre, to about 4050 for summer 2001, and 2650 for winter 2002 in suburban areas. Traffic emissions values follow the same variation.

As expected, the mean of NO_2 concentrations is lower in summer (18.0 µg/m^3) and higher in winter: 23.8 µg/m^3 in 2001 and 20.1 µg/m^3 in 2002. Is the concentration really higher in winter 2001, or is it a consequence of the preferential location of the samples? In fact, for the three samples common to both winters (which present rather high NO_2

levels), concentrations are lower in winter 2001 than in winter 2002, with average values of 25.5 and 27.7 μg/m^3 respectively. Therefore the previous global mean concentration decrease is mainly due to the different sampling locations. The scatter diagrams between winter concentration and auxiliary information strengthen this hypothesis (Fig. 2): the regression of concentration on population density does not show a systematic variation between the two winters at same population density value.

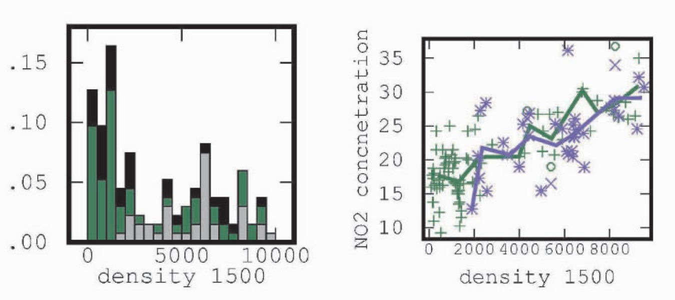


Figure 2. Population density on 1500m radius circles around the samples. Histogram: light grey, informed for winter 2001; black, additional summer 2001 data; medium gray, additional winter 2002 data. Scatter diagrams of winter concentrations versus population density, and associated empirical regression curves: * winter 2001, + winter 2002. The three common samples are respectively indicated with x and o.

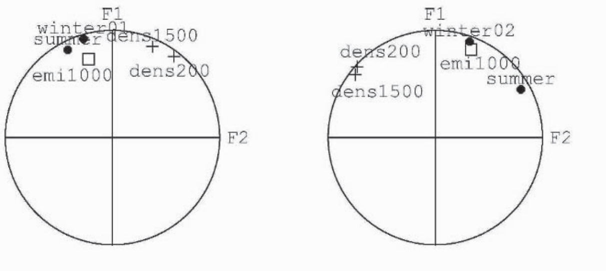


Figure 3. Principal component analysis of seasonal concentrations and auxiliary variables. Correlation circles of the first factors. Left: winter 2001 and summer 2001, right summer 2001 and winter 2002. ●: seasonal concentrations; □: NO_X emissions inventory; +: population density. The first factor represents respectively 67% and 49% of total variance.

	number of data	density200	density1000	density1500	emission1000
winter 2001	31 (36)	0.42	0.62	0.60	0.60
summer 2001	42 (57)	0.24	0.24	0.20	0.47
winter 2002	67 (88)	0.72	0.76	0.78	0.60

Table 1. Correlation coefficients between NO_2 seasonal concentrations and auxiliary variables with associated number of data. In parentheses, total number of concentration measurements. Some auxiliary information is missing.

Two Principal Component Analyses were performed on auxiliary information, summer concentration, and respectively winter 2001 or winter 2002 concentrations, keeping only winter 2001 sampling points for the first one, and winter 2002 sampling points for the second one (Fig. 3). Population densities on different supports being highly correlated, only values for radius 200m and 1500m are retained. The relationships between concentration and auxiliary information slightly differ for the two data sets: for the first one, all the variables are rather well correlated, whereas for the second one the winter concentrations are mainly correlated with road emissions.

Table 1 (calculated on the whole data set for each season) confirms that the correlation between NO_2 concentration and population density increases in winter, reflecting the influence of heating. The lower correlation in winter 2001 than in winter 2002 can result from the reduced range of the associated density values or reflects the influence of other emitters in the agglomeration centre. This correlation slightly varies with the support on which population density is given. The correlation of concentration and road traffic remains identical for the two different winter areas, and is lower in summer.

Thus the relationship between concentrations and auxiliary information depends on the local characteristics of the area. It is then necessary to check the validity of the variographic model before extending it to wider zones. An inadequate model can lead to nonsense results, as negative estimated concentration for example.

3.1.3 Relationships between winter and summer concentrations.

The correlation between winter and summer concentrations (Figure 4) is high (0.82 for winter and summer 2001, and 0.72 for summer 2001 and winter 2002) and the scatter diagram almost linear. Winter 2001 concentrations are then more correlated with summer ones, and summer concentrations much more correlated with winter ones, than with auxiliary information. For winter 2002 concentrations, the correlation coefficient is slightly higher with population density than with summer concentrations.

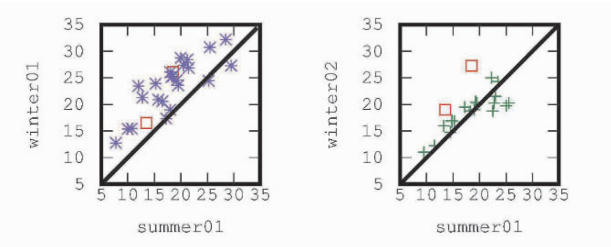


Figure 4. Scatter diagrams of winter versus summer concentrations in Montpellier. The squares denote the two samples common for the three seasons.

In conclusion, in spite of differences between the two winters and the sampled areas, the detailed exploratory data analysis shows:

- the presence of the expected relationships between NO_2 concentrations and auxiliary variables depicting the urban context,
- the importance of the local context for quantifying these relations. For example, winter concentrations are better “explained” by the population density in the suburban areas,
- the lower correlation of concentration and auxiliary variables in summer,
- the high correlation level between seasonal concentrations.

3.2 MEASUREMENTS DURING TWO SEASONS IN MULHOUSE

3.2.1 Context and sampling

Mulhouse is an old industrial agglomeration of more than 110 000 inhabitants, located near Germany and Switzerland in the south of Alsace plain (north-east of France), under continental climate. 75 urban or peri-urban sites have been monitored for NO_2 measurements during three fortnights in winter and in summer 2001. Because of technical problems, 62 seasonal measurements are available in winter, 59 in summer, and only 50 for the "yearly mean", the sampling time representing about 25% of the year (Figure 5).

The land use is given on a grid with 200m-resolution grid. Among the available land use classes, only the "dense building" is retained. Population density and NO_x emission inventory, including road traffic and industry, are given on a 1km resolution grid.

3.2.2 Exploratory analysis

The spatial correlation statistics are indicative, the sampling sites being preferentially located towards the center of the agglomeration. For the "yearly" sites, the winter average, 28.4 µg/m^3, is strongly higher than the summer one, 16.2 µg/m^3

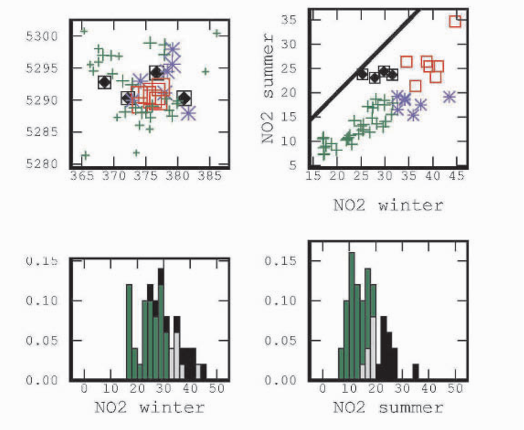


Figure 5. Mulhouse. Location (top left), scatter diagram (top right) and histograms (bottom) of seasonal NO_2 measurements. High concentrations (in black for the histograms) are marked as squares on the map and the scatter diagram, and intermediate concentrations (light grey) as stars. The stations marked with ◆ have a high summer concentration and an intermediate winter one. Distances are in km on the map.

The associated standard deviations, respectively 7.2 and 6.0 µg/m^3, show that the variability increases with concentration, whereas the relative variability, given by the dispersion coefficient (the ratio of the standard deviation to the mean) is higher in summer than in winter (respectively 0.37 and 0.25).

Low concentrations are located at the same sampling sites in winter and in summer (Fig. 5), mainly around the agglomeration. High summer concentrations correspond either to winter concentrations higher than 35µg/m^3, or to intermediate winter concentrations located in the close suburb, near important traffic infrastructures. Apart from some of those high seasonal values, the scatter diagram between winter and summer concentration is linear, with high correlation coefficient (0.82).

3.2.3 Relations with auxiliary information

Taking the translated logarithm of the auxiliary variables ($\log\left(1+\frac{z}{m}\right)$, z being the variable and m a normative factor, for example the mean) is an easy and robust way to linearise the relationships with concentration, as shown in the scatter diagrams (Figure 6). This linearity will be useful for external drift (co)kriging, which assumes a local linear relationship between the main variables and the auxiliary information.

The contribution of heating in winter, and the consequences of summer road traffic explain the different location of the high seasonal concentrations. Indeed, the correlation coefficients are similar between winter concentration and the three auxiliary variables, whereas for summer concentrations, the correlation increases with the NO_x emissions and decreases with the dense building land use (Figure 6 and Table 2). Note that all these coefficients are lower than the one between seasonal concentrations.

	Dense building	NO_x	Population density
winter	0.69	0.68	0.69
summer	0.57	0.73	0.67

Table 2. Correlation coefficients between the NO_2 seasonal concentrations and the translated logarithm of auxiliary variables.

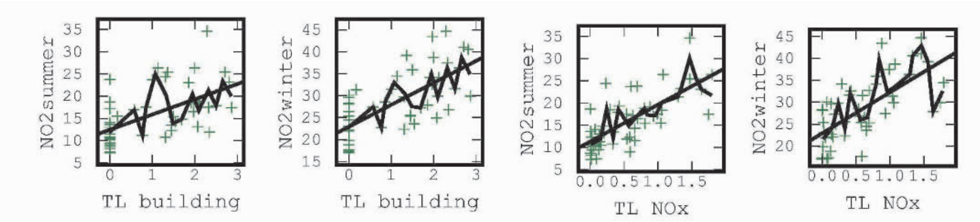


Figure 6. Scatter diagrams of NO_2 seasonal concentrations and auxiliary variables (TL for translated logarithm).

3.2.4 Conclusions

Despite the important differences of climatic and geographical context between the two agglomerations, some common conclusions can be drawn, mainly:

- the high correlation level between winter and summer NO_2 concentrations,
- some high correlation between winter concentration and auxiliary variables.

When looked in detail, the relationships between concentrations and auxiliary variables are different. The correlation with population density, similar in both seasons for Mulhouse is lower in summer for Montpellier.

Further estimations are mainly done by external drift kriging or cokriging with moving neighbourhood. The corresponding residuals being not available, the variograms are indirectly fitted by cross validation. We will then not present here the variograms of concentrations.

4 Estimation of yearly NO_2 concentration at Mulhouse

As mentioned before, each seasonal data value is approximated by the average of three fortnight measurements, and the yearly concentration by their mean.
We compare the following estimations of the yearly concentration:

- kriging with external drift, from the annual measurements only,
- average of the two seasonal kriging with external drift estimations,
- cokriging with external drift from the seasonal measurements

4.1 COKRIGING OF THE YEARLY MEAN

The non bias conditions for external drift cokriging of the annual mean $\frac{1}{2}\left(Z_1(x)+Z_2(x)\right)$ from seasonal data $Z_i, i=1,2$ are easily obtained. Setting $Z_i(x)=R_i(x)+\sum_{\ell} a_{\ell}^{i} f_i^{\ell}(x)$, where f_i^{ℓ} are the auxiliary variables (eventually after transformation such as log translation), and a_{ℓ}^{i} the unknown deterministic coefficients, the estimator is written: $\frac{1}{2}\left(Z_1(x)+Z_2(x)\right)^{*}=\sum_{\alpha}\lambda_x^{\alpha} Z_1(x_{\alpha})+\sum_{\beta}\kappa_x^{\beta} Z_2(x_{\beta})$. Assuming the a_{ℓ}^{i} without relationships between them and between winter and summer values (i.e. for $i=1,2$), the non bias condition $E\left[\left(\frac{Z_1(x)+Z_2(x)}{2}\right)-\left(\frac{Z_1(x)+Z_2(x)}{2}\right)^{*}\right]=0$ leads to $\forall\ell, \sum_{\alpha}\lambda_x^{\alpha} f_1^{\ell}(x_{\alpha})=\frac{1}{2} f_1^{\ell}(x)$ and $\forall\ell, \sum_{\alpha}\kappa_x^{\beta} f_2^{\ell}(x_{\beta})=\frac{1}{2} f_2^{\ell}(x)$. In practice, the function f_i^{1} is taken as constant, and the two conditions for $\ell=1$, also valid for cokriging with unknown means and without external drift, simply imply that the sum of the weights relative to each measurement period is equal to $\frac{1}{2}$.
It is not necessary to use the same auxiliary variables as drift for both seasons.

4.2 CROSS VALIDATION RESULTS

The classical cross-validation method is used to quantify the improvement of precision brought by auxiliary variables used as external drift or by the cokriging from the seasonal concentrations (Table 3). When available, the other seasonal data is retained at test point for cokriging. Different criteria (correlation coefficient between cross validation estimation and data, experimental variance of estimation errors, relative variance i.e. variance of the ratio of the estimation error to the data) indicate the same rank for the quality of the estimation, except for the kriging of annual concentration, with slightly different scores when comparing two sets of external drift. The main results are the following:

- As for seasonal estimation (not shown), different auxiliary variables used as external drift give almost the same cross validation result. In practice, other criteria should be considered, mainly the resulting maps, to choose the most suited, following the practical knowledge of the air pollution phenomenon (Gallois, 2004).

- Using only the 50 yearly measurements considerably reduces the precision, compared to the average of both seasonal kriging, which takes into account all seasonal measurements.

- Because of the high correlation between seasonal concentrations, cokriging drastically improves the cross validation results. As a seasonal measurement is not available on each cell of the estimation grid, this last cross validation result could be regarded as too optimistic. In fact, it has some very interesting practical consequences, for example to optimize measurement campaigns.

model	r (Z,Z*)	Var error	Var relative error
K ED : Population, dense building	.81	12.5	0.032
K ED : NO_x	.82	11.8	0.035
Mean of seasonal K ED	.85	10.2	0.030
CK ED from seasonal data	.95	3.4	0.011

Table 3. Cross validation results for “yearly” concentration. Kriging (K) and cokriging (CK) with external drift. 49 test points, because of moving neighborhood. r denotes the correlation coefficient between estimated and measured values and Var. the variance.

4.3 OPTIMIZING MEASUREMENT CAMPAIGNS

Because of the high correlation between seasonal concentrations, sampling the same sites during both seasons is partly redundant. Keeping the same number of data, sampling different sites in winter and summer would allow increasing the number of measured sites, so as to improve the estimation. Another possibility consists in removing one of the seasonal measurements, to decrease the sampling cost.

To check the feasibility of a sparse sampling, 30% of the data are removed, letting 30% of the sites sampled only in winter, 30% only in summer, and the remaining 40% being sampled during both seasons. The data to be removed are randomly drawn in space. Keeping some sites common to both seasons is necessary to model the spatial structure, in particular the cross variogram. To evaluate the actual influence of the loss of information, a cross validation of the cokriging is realised by suppressing and re-estimating successively a quarter of the stations. The model used (variograms, external drift…) is first the one drawn with all the information.

The experimental mean quadratic error on the yearly concentration is 11.7 for the reduced sampling, and 10.1 for the whole one (on exactly the same yearly sites). The precision of the estimation is only slightly reduced compared to the saving of 30% of measurements. Moreover, cokriging maps are very similar, whether computed with complete or reduced data. Two other reduced samplings were tested in the same way, changing the sets of removed data. They gave roughly the same results.

For more consistency, the model should be drawn from the reduced sampling. Therefore a new variogram is now fitted using the 70% retained data, a little different from the previous model. The new mean quadratic error on the yearly concentration becomes 10.0: the estimation seems as good as when using all measurements. In fact, re-drawing the model allows making it more appropriate to the reduced sampling. With this new model, the estimated map differs from that calculated using all measurements (Fig. 7), but the major differences are located on the edge of the map, where estimations are unreliable. The conclusion remains the same if another reduced data set is used.

As a conclusion, the high correlation between seasonal concentrations allows to significantly reduce sampling costs, without diminishing the precision level. In practice, it is necessary to keep enough samples common to both seasons to verify the validity of the multivariate variogram model, distributing them regularly not only in the geographic space but also in the space of auxiliary variables.

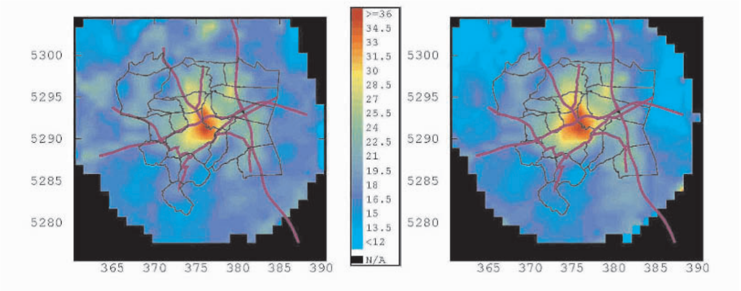


Figure 7. Cokriging of the yearly concentration from seasonal measurements, using all (left) or 70% of the data (right).

Acknowledgements

We would like to sincerely acknowledge Gilles PERRON from ASPA (Association pour la Surveillance et l'Étude de la Pollution Atmosphérique en Alsace).

References

Bobbia M., Mietlicki F., Roth C., 2000. *Surveillance de la qualité de l'air par cartographie : l'approche géostatistique*. Poster INRETS 2000, 5-8 juin, Avignon, France

Bobbia M., Pernelet V., Roth C., 2001. L'intégration des informations indirectes à la cartographie géostatistique des polluants. *Pollution Atmosphérique* n° 170 - Avril-Juin.

Cardenas G., Malherbe L., 2002. Application des méthodes géostatistiques à la cartographie de la qualité de l'air, *STIC et Environnement*, Rouen, 19-20 juin 2003.

Casado L.S., Rouhani S., Cardelino C.A., Ferrier A.J., 1994. Geostatistical analysis and visualization of hourly ozone data. *Atmospheric Environment 28*, n°12, 2105-2118.

Gallois D. 2004. *Optimisation des estimations spatio-temporelles par des méthodes géostatistiques.* Rapport de stage ASPA. Ecole des Mines de Paris.

Lefohn A.S., Knudsen H.P., McEvoy L.R., 1988. The use of kriging to estimate monthly ozone exposure parameters for the Southern United States. *Environmental Pollution*, 53, 27-42.

Liu S. L.-J., Rossini A.J., 1996. Use of kriging models to predict 12-hour mean ozone concentrations in metropolitan Toronto - A pilot study. *Environment International*, N° 6, 667-692.

Matheron G. 1970. La théorie des variables régionalisées, et ses applications. *Les cahiers du centre de morphologie mathématique de Fontainebleau*. Ecole des Mines de Paris.

Nikiforov S.V., Aggarwal M., Nadas A., Kinney P.L., 1998. Methods for spatial interpolation of long-term ozone concentrations. *Journal of Exposure Analysis and Environmental Epidemiology*, vol. 8, n°4, 465-481.

Phillips D.L., Lee E. H., Herstom A. A., Hogsett W. E., Tingey D.T., 1997. Use of auxiliary data for spatial interpolation of ozone exposure in southearn forests. *Environmetrics*, 8, 43-61.

Tayanç M., 2000. *An assessment of spatial and temporal variation of sulfure dioxide levels over Istanbul, Turkey*. Environmental pollution, 107, 61-69

A STEP BY STEP GUIDE TO BI-GAUSSIAN DISJUNCTIVE KRIGING

JULIÁN M. ORTIZ
Department of Mining Engineering, University of Chile
Av. Tupper 2069, Santiago, Chile, 837-0451

BORA OZ and CLAYTON V. DEUTSCH
Department of Civil & Environmental Engineering, 220 CEB,
University of Alberta, Canada, T6G 2G7

Abstract. The Disjunctive Kriging formalism has been implemented for a number of tasks in geostatistics. Despite the advantages of this formalism, application has been hindered by complex presentations and the lack of simple code. This paper goes through the steps to perform Disjunctive Kriging in a simple case. The global stationary distribution of the variable under consideration is fit by Hermite polynomials. The coefficients of this polynomial expansion fully define the relationship between the original values and their normal score transforms. Disjunctive Kriging amounts to using simple kriging to estimate the polynomial values at unsampled locations. The estimate of the variable is built by linearly combining the estimated polynomial values, weighted by the coefficients of fitting of the global distribution. These estimated values completely define the local distribution of uncertainty. It is straightforward to implement this formalism in computer code; this paper attempts to provide a clear exposition of the theoretical details for confident application and future development.

1 Introduction

Disjunctive Kriging (DK) has been available for more than 25 years; however the seemingly complex theory makes it unappealing for most practitioners. DK is a technique that provides advantages in many applications. It can be used to estimate the value of any function of the variable of interest, making it useful to assess truncated statistics for recoverable reserves. DK provides a solution space larger than the conventional kriging techniques that only rely on linear combinations of the data. DK is more practical than the conditional expectation, since it only requires knowledge of the bivariate law, instead of the full multivariate probability law of the data locations and locations being estimated (Maréchal, 1976; Matheron, 1973, 1976a, 1976b; Rivoirard, 1994). The theoretical basis of DK is sound, internally consistent, and has been extensively developed and expanded, among geostatisticians (Armstrong and Matheron, 1986; Emery, 2002; Maréchal, 1984; Matheron, 1974, 1984). In practice, those developments have not been applied to their full potential. DK has been applied mainly with the use of Hermite polynomials and the bivariate Gaussian assumption (Guibal and Remacre, 1984; Webster and Oliver, 1989). Still, relatively few practitioners have mastered DK. The discomfort of many practitioners is due in part to the difficult

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 1097–1102.

literature focussed on theory rather than applications. This work aims to present DK in a rigorous manner, with greater focus on its practical aspects.

We start by presenting some background on Hermite polynomials, the bivariate Gaussian assumption, and then introduce DK and its implementation steps. More extensive theory can be found in Chilès and Delfiner (1999) and Rivoirard (1994).

2 Hermite Polynomials

Before getting into DK, we need to define and review some of the properties of Hermite polynomials. This family of polynomials is important because it will help us parameterize Gaussian conditional distributions later on. Hermite polynomials are defined by Rodrigues' formula:

$$H_n(y) = \frac{1}{\sqrt{n!} \cdot g(y)} \cdot \frac{d^n g(y)}{dy^n} \qquad \forall n \geq 0$$

where n is the degree of the polynomial, $\sqrt{n!}$ is a normalization factor, y is a Gaussian or normal value, and $g(y)$ is the standard Gaussian probability distribution function (pdf) defined by $g(y) = \left(1/\sqrt{2\pi}\right) \cdot e^{-y^2/2}$. For a given value of y the polynomial of degree n, $H_n(y)$, can easily be calculated. A recursive expression, useful for computer implementation, exists to calculate polynomials of higher orders:

$$H_0(y) = 1 \qquad H_1(y) = -y \qquad H_2(y) = (y^2 - 1)/\sqrt{2}$$

$$H_{n+1}(y) = -\frac{1}{\sqrt{n+1}} \cdot y \cdot H_n(y) - \sqrt{\frac{n}{n+1}} \cdot H_{n-1}(y) \qquad \forall n \geq 1$$

These polynomials have the following properties: (1) Their means are 0, except for the polynomial of degree 0, which has a mean of 1; (2) Their variances are 1, except again for the polynomial of order 0 which is constant and therefore its variance is 0; and (3) the covariance between $H_n(Y)$ and $H_p(Y)$ is 0 if $n \neq p$. This property is known as *orthogonality* and can be understood in the same manner as the factors and principal components in multivariate statistical analysis; they correspond to uncorrelated components of a function of Y. Of course, if $n = p$ the covariance becomes the variance of $H_n(Y)$. A covariance of zero is sufficient for full independence if the bivariate distribution is Gaussian.

Hermite polynomials form an orthonormal basis with respect to the standard normal distribution, other polynomials families can be considered if a different transformation of the original variable is performed (Chilès and Delfiner, 1999).

2.1 BIVARIATE GAUSSIAN ASSUMPTION

Consider the variable Y distributed in space. We can define the random function model $\{Y(\mathbf{u}), \mathbf{u} \in \text{Domain}\}$, where $\mathbf{u}$ is a location vector in the three-dimensional space.

Taking a pair of random variables $Y(\mathbf{u})$ and $Y(\mathbf{u}+\mathbf{h})$ considered stationary, we say they are standard bivariate Gaussian if:

$$\left(Y(\mathbf{u}), Y(\mathbf{u}+\mathbf{h})\right) \sim N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \qquad with \qquad \boldsymbol{\mu} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \qquad \boldsymbol{\Sigma} = \begin{pmatrix} 1 & \rho(\mathbf{h}) \\ \rho(\mathbf{h}) & 1 \end{pmatrix}$$

Notice that these two terms, the mean vector and variance-covariance matrix, fully define the bivariate Gaussian distribution of $Y(\mathbf{u})$ and $Y(\mathbf{u}+\mathbf{h})$. The correlogram $\rho(\mathbf{h})$ gives all the structural information of the bivariate relationship.

Under this assumption, one additional property of Hermite polynomials is of interest. The covariance between polynomials of different order is always 0, and if the order is the same, it identifies the correlation raised to the polynomial's degree power, that is:

$$Cov\left\{H_n\left(Y(\mathbf{u})\right), H_p\left(Y(\mathbf{u}+\mathbf{h})\right)\right\} = \begin{cases} \left(\rho(\mathbf{h})\right)^n & \text{if } n = p \\ 0 & \text{otherwise} \end{cases}$$

The only term that is left is the covariance between polynomial values of the same degree for locations separated by a vector $\mathbf{h}$. Since $\rho(\mathbf{h}) < \rho(\mathbf{0}) = 1$, this spatial correlation tends rapidly to zero as the power n increases, that is, the structure tends to pure nugget.

2.2 FITTING A FUNCTION WITH HERMITE POLYNOMIALS

Any function with finite variance can be fitted by an infinite expansion of Hermite polynomials. The idea is to express the function of y as an infinite sum of weighted polynomial values: $f\left(y(\mathbf{u})\right) = \sum_{n=0}^{\infty} f_n \cdot H_n\left(y(\mathbf{u})\right)$

The only question that remains is how to find the coefficients $f_n, \forall n \geq 0$. This can be done by calculating the expected value of the product of the function and the polynomial of degree n:

$$\begin{aligned} E\left\{f\left(Y(\mathbf{u})\right) \cdot H_n\left(Y(\mathbf{u})\right)\right\} &= E\left\{\sum_{p=0}^{\infty} f_p \cdot H_p\left(Y(\mathbf{u})\right) \cdot H_n\left(Y(\mathbf{u})\right)\right\} \\ &= \sum_{p=0}^{\infty} f_p \cdot E\left\{H_p\left(Y(\mathbf{u})\right) \cdot H_n\left(Y(\mathbf{u})\right)\right\} = f_n \end{aligned}$$

The expected value can be taken inside the summation, since it is a linear operator and the coefficients f_p are constants. Notice that the expected value of the product of polynomials of different degrees corresponds to their covariance. The property of orthogonality comes in so that all terms where $p \neq n$ equal zero and we only have the one where $p = n$. In this case, the covariance becomes the variance that equals 1. We then obtain the expression for the coefficient f_n. It is worth noting that the coefficient of 0 degree corresponds to the mean of the function of the random variable.

The practical implementation of this expansion calls for some simplifications: the infinite expansion is truncated at a given degree P. The truncation causes some minor problems, such as generating values outside the range of the data. These values can simply be reset to a minimum or maximum value. If the number of polynomials used is large enough, these problems are of limited impact.

3 Disjunctive Kriging

Disjunctive Kriging (DK) allows the estimation of any function of $Z(\mathbf{u})$, based on a bivariate probability model. A bivariate Gaussian distribution of the normal scores of the data is almost always chosen. DK provides the solution that minimizes the estimation variance among all linear combinations of functions of one point at a time.

In simple words, DK relies on the decomposition of the variable (or a function of it) into a sum of factors. These factors are orthogonal random variables, uncorrelated with each other, and therefore the optimum estimate can be found by simple kriging each component. Consider a random variable Z and a transformed random variable Y, in general its Gaussian transform. The disjunctive kriging estimate finds the family of functions of Y that minimizes the estimation variance. Under a particular bivariate assumption, an isofactorial family of functions can be found. Under the bivariate Gaussian assumption, this family is the Hermite polynomials. However, other transformations can be done, in which cases different orthogonal polynomials must be used. They are called isofactorial families because they decompose the function of the random variable into factors that are spatially uncorrelated. Although in the general case the DK estimate is obtained by simple cokriging of the functions of different order, if these are uncorrelated from each other, just a simple kriging of the functions of the same order and their posterior linear combination suffices to obtain the best estimate.

The DK estimate is presented next under the bivariate Gaussian assumption using Hermite polynomials:

$$[f(Y(\mathbf{u}))]^{DK} = \sum_{p=0}^{\infty} f_p \cdot [H_p(Y(\mathbf{u}))]^{SK}$$

The expansion is generally truncated at a degree P, usually under 100. To calculate the DK estimate, the normal score transformation of the data is necessary. Then, the spatial covariance of the transformed variable $\rho(\mathbf{h})$ is calculated and modelled (it is the correlogram, since it has unit variance). The Hermite polynomials are computed for all the transformed data up to a degree P. Finally, the coefficients of the Hermitian expansion can be calculated. Simple kriging is performed P times. The estimate of the Hermite polynomial at an unsampled location $\mathbf{u}$ is calculated as:

$$[H_p(y(\mathbf{u}))]^{SK} = \sum_{i=1}^{n(\mathbf{u})} \lambda_{p,i} \cdot H_p(y(\mathbf{u}_i)) \qquad \forall p > 0$$

where $\lambda_{p,i}$ is the simple kriging weight for datum $y(\mathbf{u}_i)$ and the degree p; $n(\mathbf{u})$ is the number of samples found in the search neighborhood used for kriging. Notice that the term for the mean is not present, since the mean value of the Hermite polynomial is 0, for all $p > 0$. Also, note that the SK estimate for the polynomial of degree 0 is 1.

The weights are obtained by solving the following system of equations:

$$\begin{bmatrix} (\rho_{1,1})^p & \cdots & (\rho_{1,n(\mathbf{u})})^p \\ \vdots & \ddots & \vdots \\ (\rho_{n(\mathbf{u}),1})^p & \cdots & (\rho_{n(\mathbf{u}),n(\mathbf{u})})^p \end{bmatrix} \cdot \begin{bmatrix} \lambda_{p,1} \\ \vdots \\ \lambda_{p,n(\mathbf{u})} \end{bmatrix} = \begin{bmatrix} (\rho_{1,0})^p \\ \vdots \\ (\rho_{n(\mathbf{u}),0})^p \end{bmatrix}$$

We can now rewrite the DK estimate as:

$$[f(Y(\mathbf{u}))]^{DK} = \sum_{p=0}^{\infty} f_p \cdot \left[\sum_{i=1}^{n(\mathbf{u}_0)} \lambda_{p,i} \cdot H_p(y(\mathbf{u}_i))\right]$$

4 Implementing DK

Implementation of disjunctive kriging requires the following steps and considerations:

1. The original data $z(\mathbf{u}_1),...,z(\mathbf{u}_N)$ must be transformed to normal scores $y(\mathbf{u}_1),...,y(\mathbf{u}_N)$.
2. The Hermite polynomials of each data are calculated up to a degree P : $H_1(y(\mathbf{u}_i)),...,H_P(y(\mathbf{u}_i)),\ \forall i = 1,...,N$.
3. The coefficients f_p, $p = 1,...,P$ are calculated. Notice that the function $f(Y(\mathbf{u}))$ may simply correspond to the inverse transformation function from Z to Y , or it may be a more complex function of Y . The coefficients are calculated as a discrete sum. For example, if the function is the inverse transformation to normal scores, then:

$$\begin{aligned} f_n &= E\{f(Y(\mathbf{u}))\cdot H_n(Y(\mathbf{u}))\} = \int f(y)\cdot H_n(y)\cdot g(y)\cdot dy \\ &= \sum_{i=1}^{N} \int_{y(\mathbf{u}_i)}^{y(\mathbf{u}_{i+1})} z(\mathbf{u}_i)\cdot H_n(y)\cdot g(y)\cdot dy \\ &= \sum_{i=2}^{N} (z(\mathbf{u}_{i-1}) - z(\mathbf{u}_i))\cdot \frac{1}{\sqrt{p}}\cdot H_{n-1}(y(\mathbf{u}_i))\cdot g(y(\mathbf{u}_i)) \end{aligned}$$

   If the function is the probability for the node to be below a threshold, that is, its indicator function, then:

$$\begin{aligned} f_n &= E\{f(Y(\mathbf{u}))\cdot H_n(Y(\mathbf{u}))\} = \int I_Y(u; y_c)\cdot H_n(y)\cdot g(y)\cdot dy \\ &= \int_{-\infty}^{y_C} H_n(y)\cdot g(y)\cdot dy = \frac{1}{\sqrt{p}}\cdot H_{n-1}(y_C)\cdot g(y_C) \end{aligned}$$

   The coefficients can be calculated in the same manner for any function of Y .
4. The variogram of normal scores must be calculated and modelled. This provides us with the correlogram, which fully defines the spatial continuity for polynomials of different degree.
5. At every location to be estimated, P simple kriging systems are solved, one for each degree of the polynomials, using the covariance (correlation) function modelled for the normal scores raised to the power p of the degree of the polynomial being estimated. These systems provide a set of estimated Hermite polynomials for the unsampled location, which are then linearly combined using the coefficients calculated in Step 3:

$$\left[f\left(Y(\mathbf{u})\right)\right]^{DK} = \sum_{p=0}^{P} f_p \cdot \left[H_p\left(Y(\mathbf{u})\right)\right]^{SK}$$

5 Conclusions

This paper presents the methodology to estimate the value of a regionalized variable at an unsampled location by Disjunctive Kriging, under the bivariate Gaussian assumption. The use of the Hermite polynomials as an isofactorial family was discussed and the most fundamental equations were presented. The methodology presented here could be extended to other transformations using different isofactorial families.

Acknowledgements

This paper was prepared while the principal author was a Ph.D. student at the Centre of Computational Geostatistics (CCG) in Edmonton, Alberta. He would like to thank the CCG for funding and acknowledge the Department of Mining Engineering at University of Chile and its industry sponsor Codelco Chile for supporting this research.

References

Armstrong, M., and Matheron, G., Disjunctive kriging revisited (Parts I and II). *Mathematical Geology*, 18(8):711-742, 1986.

Chilès, J. P., and Delfiner, P., *Geostatistics: Modeling spatial uncertainty*, Wiley, New York, 696 p, 1999.

Emery, X., Conditional simulation of non-Gaussian random functions. *Mathematical Geology*, 34(1):79-100, 2002.

Guibal, D., and Remacre, A., Local estimation of the recoverable reserves: Comparing various methods with the reality of a porphyry copper deposit, *in* Geostatistics for Natural Resources Characterization, G. Verly, M. David, A. G. Journel, and A. Maréchal, editors, Reidel, Dordrecht, Holland, vol. 1, pp. 435-448, 1984.

Maréchal, A., The practice of transfer functions: Numerical methods and their applications, *in* Advanced Geostatistics in the Mining Industry, M. Guarascio, M. David, and C. Huijbregts, editors, Reidel, Dordrecht, Holland, pp. 253-276, 1976.

Maréchal, A., Recovery estimation: A review of models and methods, *in* Geostatistics for Natural Resources Characterization, G. Verly, M. David, A. G. Journel, and A. Maréchal, editors, Reidel, Dordrecht, Holland, vol. 1, pp. 385-420, 1984.

Matheron, G., Le krigeage disjonctif. Internal note N-360, Centre de Géostatistique, Fontainebleau, 40 p., 1973.

Matheron, G., Les fonctions de transfert des petits panneaux. Internal note N-395, Centre de Géostatistique, Fontainebleau, 73 p., 1974.

Matheron, G., A simple substitute for conditional expectation: the disjunctive kriging, *in* Advanced Geostatistics in the Mining Industry, M. Guarascio, M. David, and C. Huijbregts, editors, Reidel, Dordrecht, Holland, pp. 221-236, 1976a.

Matheron, G., Forecasting block grade distributions: The transfer function, *in* Advanced Geostatistics in the Mining Industry, M. Guarascio, M. David, and C. Huijbregts, editors, Reidel, Dordrecht, Holland, pp. 237-251, 1976b.

Matheron, G., Isofactorial models and change of support, *in* Geostatistics for Natural Resources Characterization, G. Verly, M. David, A. G. Journel, and A. Maréchal, editors, Reidel, Dordrecht, Holland, vol. 1, pp. 449-467, 1984.

Rivoirard, J., *Introduction to Disjunctive Kriging and Non-Linear Geostatistics*, Oxford University Press, 181 p., New York, 1994.

Webster, R., and Oliver, M. A., Disjunctive kriging in agriculture, *in* Geostatistics, M. Armstrong, editor, vol. 1, pp. 421-432, Kluwer, 1989.

ASSESSING THE POWER OF ZONES OF ABRUPT CHANGE DETECTION TEST

EDITH GABRIEL and DENIS ALLARD
Institut National de la Recherche Agronomique, Unité de biométrie
Domaine Saint-Paul, site Agroparc, 84914 Avignon, Cedex 9, France

Abstract. We assess the power of a test used to detect Zones of Abrupt Change (ZACs) in spatial data, in the center of a moving window. Mapping the power allows to identify zones where ZACs may be detected.

1 Introduction

Allard *et al.* (2004) and Gabriel *et al.* (2004) proposed a method for detecting Zones of Abrupt Change (ZACs) for a random field, $Z(\cdot)$, defined on $\mathcal{D} \subset \mathbf{R}^2$. ZACs are defined as a discontinuity or a sharp variation of the local mean. The method is basically a test based on the estimated local gradient, where the null hypothesis $H_0(x)$: "$E[Z(y)]$ is constant for all y in a neighborhood around x" is tested against the alternative $H_1(x)$: "x belongs to Γ", where Γ is a curve on which the expectation of the random field is discontinuous.

In this paper, the power of this test is investigated. For calculating a power, the alternative must be fully specified: at a point x, we will assume that Γ has a regular shape and can be locally approximated by a line with unknown direction containing x. The punctual power is then computed for each point x of $\mathcal{D}$. The local power of a small window $\mathcal{F}$ centered on x is also considered. The difference between the punctual power and the local power is that several local test statistics are used for computing the local power. These local test statistics are not independent and computing the local power must take into account these dependencies. Mapping punctual or local power shows clearly that the power is not constant on the domain. Zones with low power indicate that the local sampling pattern is not appropriate for estimating ZACs, in particular because the local sampling density is too low.

This paper is organized as follows. The method for detecting ZACs is described in section 2 and the assessment of the power is presented in section 3. In section 4, theoretical results are illustrated on soil data of an agricultural field on which previous analysis has shown sharp transitions between different zones. It is shown how the sampling density directly affects the possibility of detecting the ZACs.

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 1103–1108.

2 Detecting Zones of Abrupt change

In this section we only present the main features of the method for detecting ZACs and we refer to Allard *et al.* (2004) and Gabriel *et al.* (2004) for more details.

$Z(\cdot)$ is assumed to be a Gaussian random field. We assume the covariance function of $Z(\cdot)$ exists, is stationary :

$$\forall x, y \in \mathcal{D},\ \operatorname{Cov}(Z(x), Z(y)) = C_Z(x-y),$$

and regular enough. Assuming second order stationarity with known expectation m, we can consider without loss of generality that $m = 0$. The interpolated local gradient is according to Chilès and Delfiner (1999),

$$W^*(x) = \nabla Z^*(x) = (\nabla C(x))'\mathbf{C}^{-1}Z, \tag{1}$$

where $Z^*(x) = C'(x)\mathbf{C}^{-1}Z$ the simple kriging of $Z(\cdot)$ at an unsampled location x, $C(x) = (C_Z(x-x_1), \ldots, C_Z(x-x_n))^{'}$, $\mathbf{C} = E[ZZ^{'}]$ is the covariance matrix with elements $\mathbf{C}_{[ij]} = C_Z(x_i - x_j)$ and $Z = (Z(x_1), \ldots, Z(x_n))^{'}$ is the sample vector. In general, the expectation is unknown and the spatial optimal predictor is the ordinary kriging (Cressie, 1993). In this case, we replace $\mathbf{C}^{-1}$ by $\mathbf{K}^{-1} = (\mathbf{C}^{-1} - \mathbf{C}^{-1}\mathbf{1}\mathbf{1}'\mathbf{C}^{-1})/\mathbf{1}'\mathbf{C}^{-1}\mathbf{1}$ in (1) and the method remains formally identical.

Detecting ZACs consists in testing the null hypothesis H_0: "$E[Z(x)] = m$ for all $x \in \mathcal{D}$" versus H_1: "the discontinuities of $E[Z(\cdot)]$ define some set of curves Γ". To discontinuity jumps correspond local high gradient of the interpolated variable, denoted Z^*. This global test is the aggregation of local tests.

2.1 LOCAL DETECTION OF A DISCONTINUITY

First, we test $H_0(x)$: "$E[Z(y)] = 0$, locally, for all y in a neighborhood around x", versus the alternative $H_1(x)$: "$x \in \Gamma$". Let us denote $\mathbf{\Sigma}(x)$ the covariance matrix of the interpolated gradient. Then, under $H_0(x)$ the statistic

$$T(x) = W^*(x)^{'}\mathbf{\Sigma}(x)^{-1}W^*(x)$$

has a $\chi^2(2)$ distribution. The local null hypothesis is rejected if $T(x) \geq t_\alpha$, where t_α is the $(1-\alpha)$-quantile of the $\chi^2(2)$ distribution.

This procedure can be applied on the nodes of a grid superimposed on the domain $\mathcal{D}$ under study, which enables us to map the regions where the local test is rejected. We define potential ZACs as: $\{x : T(x) \geq t_\alpha\}$. The size of the potential ZACs depends on α: when α decreases, ZACs become smaller; but the depth of the ZACs are linked to the sample density: high sample density leads to better detection and more precision about the location of ZACs.

If the random field is stationary, we should find a proportion of about α pixels in which $H_0(x)$ is rejected. The tests in neighboring grid nodes are not independent, so we should expect the grid nodes where $H_0(x)$ is rejected to be structured in randomly located small patches. On the contrary, if there is a discontinuity, or a sharp variation, of the expectation field, we must expect the grid nodes where $H_0(x)$ is rejected to be much more numerous and located on, or near, the discontinuity Γ.

2.2 GLOBAL TEST FOR POTENTIAL ZACS

We aggregate the local tests in the global test using the following result. Under H_0 the size of a potential ZAC, say S_{t_α}, is related to the curvature of T at the local maxima in the potential ZAC, according to

$$t_\alpha S_{t_\alpha} \det(\mathbf{\Lambda})^{1/2}/\pi \xrightarrow{\mathcal{L}} E(2), \ as \ t_\alpha \to \infty,$$

where $E(2)$ is an exponential random variable with expectation 2 and $\mathbf{\Lambda}$ is the 2×2 matrix of the curvature of T at local maximum in the potential ZAC, whose expression only depends on $C_Z(h)$ and the sampling pattern and can be found in Allard *et al.* (2004). When t_α is very large, there is a high probability that at most one potential ZAC exists on $\mathcal{D}$ under H_0 and testing on each potential ZAC is then equivalent to testing on the entire domain $\mathcal{D}$. At a global level of confidence $1-\eta$, H_0 will be rejected if $\exp\left(-t_\alpha S_{t_\alpha} \det(\mathbf{\Lambda})^{1/2}/2\pi\right) < \eta$. In this case the potential ZAC is significant and is called a ZAC.

2.3 DETERMINATION OF α AND COVARIANCE ESTIMATION

The value of the local level of confidence $1-\alpha$ that achieves the global one $1-\eta$ is a function of η, but also depends on other parameters such as the mesh of the grid on which computations are performed, the range parameter of the covariance function, the density of samples, etc. The appropriate level α is found by Monte-Carlo simulations under H_0: a series of N simulations of a Gaussian field under the null hypothesis, conditional on the same discretization, sample locations and covariance function is performed. The level $\hat{\alpha}$ corresponding to η is then defined as $\hat{\alpha} = \sup\{\alpha : M_\alpha \leq \eta N\}$, where M_α is the number of simulations (among N) with significant ZACs.

The method assumes that the covariance function is known. In practice, it must be estimated along with the potential ZACs. In a simulation study, we showed (Gabriel *et al.*, 2004) that our method is robust with respect to reasonable variations of the covariance parameters (range and parametric family). Under the alternative hypothesis, the presence of discontinuities of the expectation field implies an overestimation of the variance and consequently a loss of power of the method. To solve this difficulty, we proposed an iterative procedure, in which the covariance function, α and ZACs are estimated at each step. In the covariance estimation all pair of samples $\{Z(x_i), Z(x_j)\}$ for which the segment $[x_i, x_j]$ intersects a potential ZAC are discarded from the estimation procedure. Convergence is reached when the set of ZACs remains unchanged.

3 Power of the local test

The power of the local test corresponds to the probability to reject the null hypothesis of stationarity under the alternative of existence of a discontinuity. The problem is that the shape of the discontinuity must be specified in order to assess the poxer of the test. The method presented above leaves free the shape of the Zones of Abrupt Change. For assessing the power, we suppose that the discontinuity has a regular shape. This hypothesis allows us to approach the discontinuity by its tangent at the point where the power is assessed. Therefore, we consider the following alternative hypothesis : $E[Z(x)]$ presents a discontinuity represented by a line containing x. Under this alternative,

$$W^*(x) = W^*_{H_0}(x) + k_a(x) = (\nabla C(x))'\mathbf{C}^{-1}Z + (\nabla C(x))'\mathbf{C}^{-1}A(x),$$

where $A(x)$ is a n-vector. Its elements are $\pm a/2$ depending on which side of the discontinuity the data point is located. The power must take into account all the information contained on $\mathcal{D}$. As the local test statistics are not independent, computing the power is difficult. Hence, we will calculate the power at a point x using the information contained in the neighborhood around x. We consider x to be the center of a window $\mathcal{F}_k \subseteq \mathcal{D}$ containing $(2k+1) \times (2k+1)$ pixels. For an increasing sequence of windows, $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \ldots \subseteq \mathcal{F}_k \subseteq \ldots \subseteq \mathcal{D}$, we thus obtain an approximation of the power:

$$1 - \beta(\mathcal{F}_0) \leq 1 - \beta(\mathcal{F}_1) \leq \ldots \leq 1 - \beta(\mathcal{F}_k) \leq \ldots \leq 1 - \beta. \tag{2}$$

Because the orientation of the discontinuity is unknown, we consider a uniform orientation:

$$1 - \beta(\mathcal{F}_k) = \frac{1}{\pi}\int_0^\pi \{1 - \beta(\mathcal{F}_k(\theta))\}\, d\theta, = \frac{1}{\pi}\int_0^\pi P_{H_1(x,\theta)}[H_0(x) \text{ is rejected}] d\theta.$$

$H_1(x,\theta)$ is the existence of a linear discontinuity containing x and with angle θ.

PUNCTUAL POWER

To calculate the power at the center of the minimal window $\mathcal{F}_0 = \{x\}$, we use the decomposition of $T(x)$ as the sum of two non independent squared gaussian random fields with variance 1, $U_1(x)$ and $U_2(x)$, centered under H_0 and with expectation $\mu_i(x; a, \theta), i = 1, 2$ in presence of a discontinuity. The power on x is:

$$1 - \beta(\mathcal{F}_0) = \frac{1}{\pi}\int_0^\pi \{1 - \beta(\mathcal{F}_0(\theta))\}\, d\theta, \text{ with } 1 - \beta(\mathcal{F}_0(\theta)) = 1 - F_{\chi^2(2,\, \mu_\theta(x))}(t_\alpha),$$

where $\mu_\theta(x) = \sum_{i=1}^2 \mu_i^2(x; a, \theta)$ is the non-centrality parameter.

LOCAL POWER

Using the parametric form of the circle $U_1^2(x) + U_2^2(x) = t$:

$$U_1(x) = \sqrt{t}\cos(\omega),\ U_2(x) = \sqrt{t}\sin(\omega), \tag{3}$$

the power on x considering the information contained in a window $\mathcal{F}_k$ having N points $x_1, ..., x_N$ is:

$$1-\beta(\mathcal{F}_k)=\frac{1}{\pi}\int_0^\pi \{1 - \beta(\mathcal{F}_k(\theta))\}\, d\theta, \text{ with } 1-\beta(\mathcal{F}_k(\theta))=\frac{1}{2^N}\int_{\omega\in[0,2\pi]^N}\int_{\mathbf{t}\in[t_\alpha,\infty)^N} f_{V_\theta}(v) d\mathbf{t} d\omega,$$

where V_θ is a $2N$-gaussian vector with elements U_1 and U_2 in $x_1, ..., x_N$ under the alternative, f_{V_θ} is its density, and the parameters $\mathbf{t} = (t_{x_1}, ..., t_{x_N})'$, $\omega = (\omega_1, ..., \omega_N)'$ come from (3).

4 Application to a soil data set

The data considered here are the soil water content (mm), sampled in October 2000 and August 2001 in an agricultural field (10 ha) in Chambry, Northern France, on a pseudo-regular grid with 66 and 77 points respectively for each date (distance between nodes = 36 m). These variables are presented in the following table:

Date	Min.	1st Q.	Median	Mean	3rd Q.	Max.	σ
October 2000	136.7	232.9	257.1	270.7	313.2	412.5	49.47
August 2001	130.6	240.0	261.7	275.6	318.5	448.1	54.64

In figures 4*a* and 4*c*, the interpolation map of the variables on a 62 × 98 grid display high (resp. low) values in light (resp. dark) grey. Samples are superimposed on these images. It seems that the total moisture is lower in the Northern part of the field than in the Southern one. A zone with lower moisture is also visible in the Eastern part. The results of the analysis are shown in figures 4*b* and 4*d*. Significant ZACs are depicted in black. Two ZACs near the edge of the central part of the field appeared for both dates. For October 2000, the ZAC in the Western part is not significant. In the central part no ZAC is estimated because this area corresponds to a smooth transition. Applying the method to the soil water content for about ten dates (from March 2000 to July 2003), we showed that the ZAC in the Western part is permanent. It is only badly detected for October 2000. It can be due to the lack of sample points in the South-Eastern part of the field for this date. The assessment of the power at each pixel allows to confirm this result. The practical implementation leads to consider a square window of sidewidth $2k + 1$ pixels. The window is moved on the domain and the power is assessed for different size: $k \in \{0, 3, 5\}$ with a discontinuity $a = 3$ standard-error (respectively figures 2*a*, 2*b*, 2*c* for October 2000 and figures 2*d*, 2*e*, 2*f* for August 2001). The power is the mean of $1 - \beta(\mathcal{F}_k(\theta))$ calculated for four directions of the discontinuity. The case $k = 0$ corresponds to the punctual power. Light grey represents values near 1 and the dark one near 0.

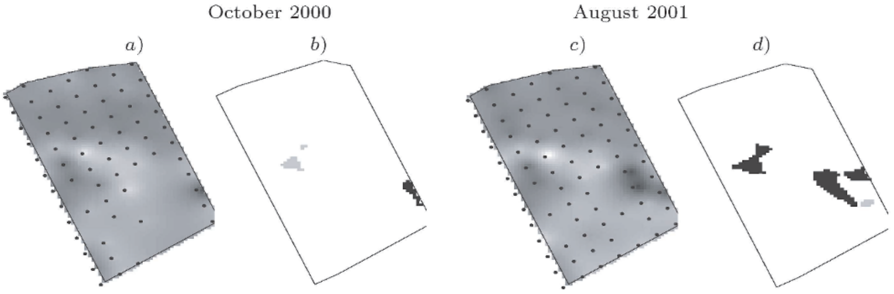


Figure 1. *a) and c): Interpolation and sampling pattern, b) and d): ZACs (significant in black).*

These results illustrate the inequalities (2): the larger the window, the higher the power. If we compare these results and the ones in figures 4*b* and 4*d*, we see

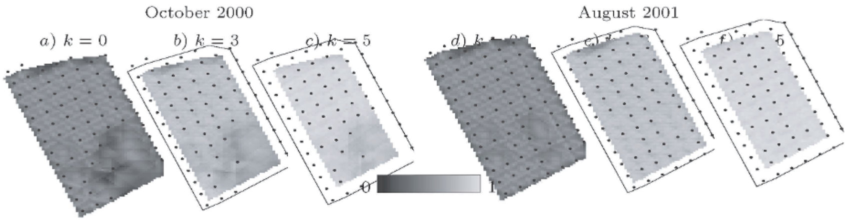


Figure 2. *Punctual ($k = 0$) and local ($k \geq 1$) power assessed in the center of a squared moving window with sidewidth $2k + 1$ pixels.*

that even if the ZACs obtained in the Eastern part of the field for August 2001 would exist in October 2000, they could not be detected. The zone with low power define the place where the local sample density is too weak for detecting potential ZACs.

5 Discussion

We approximated the power of the ZAC detection test assessed in the centre x of a moving window $\mathcal{F}$. Mapping the power allows to display the zones where the local sampling pattern is not appropriate for estimating ZACs because its local density is probably too low.

There is still an open question: for which size of the window can we consider that we have a good approximation of the power ? One might think that the window must be as large as possible to take into account all the information provided by the sample locations. For example, considering a window size equal to the practical range leads in our experience to quite large windows, probably larger than appropriate. Indeed, as the window increases the approximation of the discontinuity by its tangent becomes less acceptable. The optimal window size is thus the result of a trade-off and further research based on simulation studies are necessary.

References

Allard D., Gabriel E. and Bacro J.-N., *Estimating and testing zones of abrupt change for spatial data,* 2004 (*In revision in JASA*).

Chilès J.-P. and Delfiner P., *Geostatistics: modeling spatial uncertainty* New-York: Wiley, 1999.

Cressie N., *Statistics for spatial data*, New-York: Wiley, 1993.

Gabriel E., Allard D. and Bacro J.-N., *Detecting zones of abrupt change in soil data*, In X. Sanchez-Vila, J. Carrera and R. Froideveaux (Eds.) Proceedings of the IV International Conference on Geostastistics for Environmental Applications, Kluwer Academic Publisher, 2004, pp 437-448.

EXPERIMENTAL STUDY OF MULTIPLE-SUPPORT, MULTIPLE-POINT DEPENDENCE AND ITS MODELING

SUNDERRAJAN KRISHNAN
Department of Geological and Environmental Sciences,
Stanford University, CA-94305

Abstract

In flow related studies, data comes from widely different volume supports, each involving different non-linear, multiple-point averages. The redundancy of these data and their respective information contents with regard to a reference variable (the unknown) are evaluated using a common, large and finely gridded 2D training image (Ti). A consistent vector of data mimicking well log data, well test data and large scale seismic impedance data together with the reference flow-based effective permeability values of constant size blocks (the unknown) are first obtained from the fine gridded Ti.

The complex joint dependence between sets of these variables are investigated through conditional probabilities directly read from the Ti vector. The exactly fully conditional probability of the unknown is obtained directly from that Ti vector.

The individual single-datum conditional probabilities from different supports are combined together using the tau model (Journel, 2002). The exact tau weights are computed using knowledge of the joint distribution from the Ti-vector. These exact weights are compared with weights obtained from a calibration-based technique. It is shown that using the tau weights corrects for the severe bias resulting from algorithms which would assume the data to be conditionally independent.

Introduction

Probabilistic data integration is an important problem in many branches of earth sciences. Data coming from different supports and measurement devices inform about an unknown in varying degrees. Such information can be represented as conditional distributions of the unknown given the data. Here, we consider the problem of combining various such probabilistic data arriving from different supports. Using the tau model, we demonstrate that accounting for data redundancy is critical, and ignoring it can lead to severe bias and inconsistencies.

The data set

Figure 1 shows the reference data set. Figure 1a shows the fine support permeability distribution of size 500 x 500. This has been generated using the sequential simulation program *sgsim* (with locally varying mean) (Deutsch and Journel, 1996) using a

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 1109–1115.

variogram range of 30 and 10 in East-West (E-W) direction and North-South (N-S) respectively. Further, randomly located low permeability shales of average length 50 pixels in the E-W direction and 5 pixels in the N-S direction were superimposed on this high-perm matrix. Figure 1b has been obtained as a geometric average of this fine support data over a 11 x 11 window. This represents the support of interest, say of input into a flow simulator which predicts the transport of fluid through the geological medium. Figures 1c and 1d represent two different types of flow based well-test averages. Each was obtained by a different combination of power averages. Data shown in Figure 1c is obtained by an arithmetic average of 8 radial harmonic averages representing radial flow outwards. Data shown in Figure 1d is obtained by a harmonic

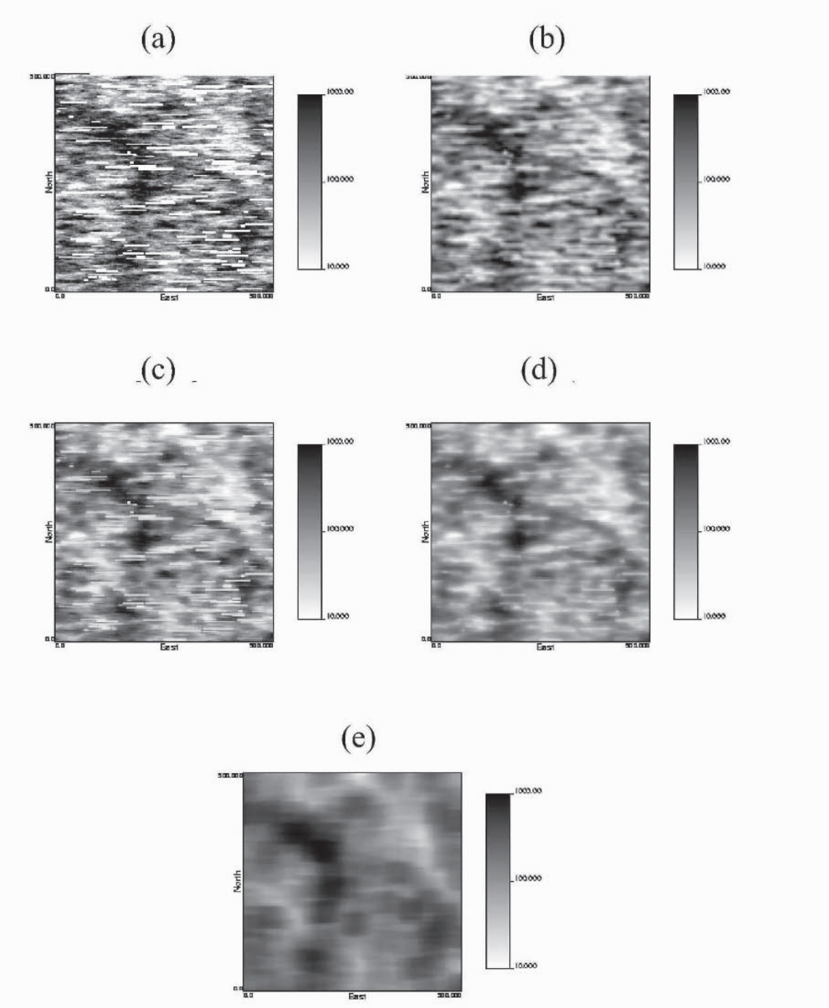


Figure 1: Vectorial training image: a) fine scale, b) 11 by 11 geometric average, c) Radial well-test, d) Annular well-test and e) 51 x 51 linear averages

average of annular geometric averages representing regions of radially outward flow. Figure 1e shows a large-scale linear average over a 51 x 51 window. This represents Data obtained from a seismic type survey.

One measure of flow-based information provided by data from any support is the multiple-point connectivity at that particular support (Krishnan and Journel, 2003). This measure of connectivity is defined as the proportion of connected strings of pixels in a particular direction (with values greater than a threshold limit). This can also be interpreted as the probability P(A) of observing a string of connected pixels of a given length. Here, we use the upper quartile as the threshold for high-valued pixels.

The lowest curve in Figure 2 is the marginal connectivity in East-West direction for the reference image of Figure 1b at the 11 x 11 support. Note that the connectivity at lag 1 is equal to the proportion 0.25 of high valued pixels and it drops steadily with increasing lag, i.e., P(A) drops from 0.25 to 0.

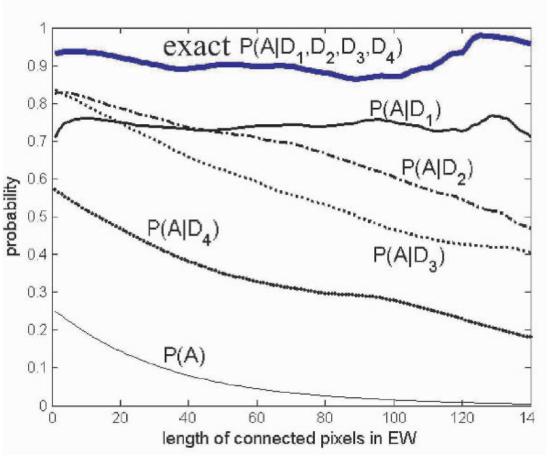


Figure 2: Marginal and conditional probabilities:

Now consider the concept of conditional connectivity. Generally, we do not have data defined at the resolution of our model, here at the 11 x 11 support. One has to infer the statistics at this support, given data from other supports. Provided we can compute the connectivity at the point support, what is the connectivity at the 11 x 11 support? The *conditional connectivity* P(A|D$_i$) denotes the probability of observing a connected string A at the 11 x 11 support, given D$_i$, i.e., a colocated set of connected pixels at another support. Here D$_1$ is the point support, D$_2$ is the first well-test, D$_3$ is the second well-test and D$_4$ is the seismic-based average. Figure 2 shows these individual conditional connectivities as well as the joint connectivity P(A|D$_1$, D$_2$, D$_3$,D$_4$). Here, note that each of the four data D$_i$ yields a probability greater than the marginal probability P(A). The overall probability P(A|D$_1$, D$_2$, D$_3$,D$_4$) is greater than all these individual *agreeing* information. This strong compounding of information implies that a simple linear averaging of all these probabilities would result in a too low estimate of the joint probability. Note that the computed connectivity values at higher lags are less reliable due to ergodic limits. A more detailed study of the multiple-support relationships between data D$_1$ through D$_4$ is performed in a recent PhD thesis (Krishnan, 2004). In order to properly combine these different individual information P(A|D$_i$), one needs a technique which accounts for redundancies between the 4 data and allows for compounding, in this case, P(A|D$_1$, D$_2$, D$_3$,D$_4$) > max{P(A|D$_i$), i=1,…,4}. Such a formula is the tau model developed extensively in the companion paper within this volume (Krishnan et al, 2004).

The Tau model

Given the marginal probability P(A) and conditional probabilities P(A|D$_i$), the tau model (Journel 2002; Bordley, 1982) gives an expression to compute the combined probability P(A|D$_1$,…,D$_n$). Define the following distances to the event A occuring (A is a connected string of a given length at the 11 x 11 support) :

$$x_0 = \frac{1 - P(A)}{P(A)},\; x_1 = \frac{1 - P(A \mid D_1)}{P(A \mid D_1)}, \ldots.,\quad x_n = \frac{1 - P(A \mid D_n)}{P(A \mid D_n)} \in [0, \infty\ ]$$

The target distance x is given by:

$$x = \frac{1 - P(A \mid D_1, \ldots, D_n)}{P(A \mid D_1, \ldots, D_n)} \in [0, \infty\ ]$$

The combined conditional probability P(A|D$_1$,…,D$_n$) is given by 1/(1+x) and it always lies in [0,1].
The tau model gives the unknown distance x as:

$$\frac{x}{x_0} = \prod_{i-1}^{n} \left(\frac{x}{x_i}\right)^{\tau_i} \qquad (1)$$

$\tau_i \in (-\infty, \infty)$. Note here, that any distance x$_i$ equal to 0 or ∞ results in a final distance x = 0 or ∞, that is, individual certainty results in global certainty. However, if there is a conflict between data, eg. x$_i$ = 0 and x$_j$ =∞, then that conflict would need to be resolved prior to using the tau model. The companion paper (Krishnan et al, 2004) develops the tau model by decomposition of the joint probability P(A|D$_1$,…,D$_n$). It expresses the tau weights as a function of the joint dependency between the datum D$_i$ and all previous data, $\overline{D}_{i-1}^{(s)} = D_1^{(s)}, \ldots, D_{i-1}^{(s)}$ taken along a specific sequence *s*:

$$\tau^{(s)}_{i,\overline{D}_{i-1}} = \frac{\log\left[\dfrac{P(D_i^{(s)} = d_i^{(s)} \mid A = a, \overline{D}_{i-1}^{(s)} = \overline{d}_{i-1}^{(s)})}{P(D_i^{(s)} = d_i^{(s)} \mid A \neq a, \overline{D}_{i-1}^{(s)} = \overline{d}_{i-1}^{(s)})}\right]}{\log\left[\dfrac{P\,(D_i^{(s)} = d_i^{(s)} \mid A = a)}{P(D_i^{(s)} = d_i^{(s)} \mid A \neq a)}\right]} \qquad (2)$$

Note that these tau weights are dependent on the specific ordering of data, i.e., D$_1$, D$_2$ would give a different set of tau weights (τ_1, τ_2) than by taking first D$_2$ followed by D$_1$. One can average the tau weights over all possible sequences, s = 1, … , S, i.e. in case of 4 data we would have 4! = 24 such sequences. That would give the averaged weights $\overline{\tau}_i$. Taking product of different tau model expressions will show that these averaged tau weights $\overline{\tau}_i$ are also exact and would retrieve the combined probability P(A|D$_1$,…,D$_n$).

It is also shown in the companion paper that assuming the data to be conditionally independent given the unknown would result in equal tau weights of 1. Note that the reverse is not necessarily true. In other words, putting the tau weights equal to 1 is not equivalent to assuming conditional independence, it is a slightly less restricting assumption.

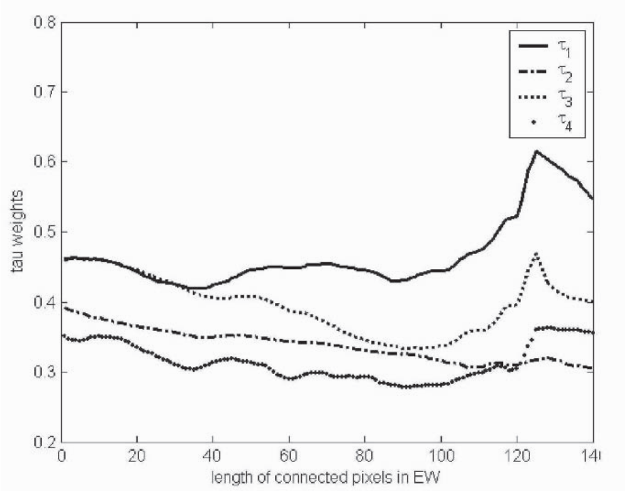


Figure 3: Exact tau weights $\overline{\tau}_i$

The companion paper suggests a calibration technique to approximate the tau weights. Instead of evaluating the joint dependency between all data, the method involves first an information content-based ranking of the data. Then, the most informative datum receives a maximum weight of 1, and any other datum D_i receives a weight restricted to be in [0,1]. This weight is a function of the conditional correlation ρ_i of datum D_i with the most informative datum D_1 :

$$\tau_1 = 1; \qquad \tau_i \;=\; 1 - (\rho_i^2)^{f(t)} \quad ; t \in \; [0,1] \;\; ; \;\; f(t) = \log(1/(1-t)) \qquad (3)$$

This method involves a calibration parameter *t* which can be computed from training information.

Evaluating the tau weights

The tau weights are used to evaluate the combined probability $P(A|D_1, D_2, D_3 ,D_4)$ from the four individual datum conditional probabilities. Here, we compute the tau weights using both the exact expression (2) and the calibration-based approximation (3). The vectorial Ti data described before is used in both cases. Figure 3 shows the exact tau weights $\overline{\tau}_i$ and Figure 4 shows the tau weights from calibration. The exact tau weights using (2), then averaged over all possible sequences, are quasi-constant and are reasonably equal one to each other. On the other hand, defining a data sequence, setting the tau weight equal to 1 for the most informative datum and restricting the others in [0,1], results in all data other than the first, receiving a much lesser weight. Note that the point support (D_1) receives maximum tau weight in Figure 3 and the largest support (D_4) receives the least weight. The approximate invariance of these exact weights (Figure 3) with lag along with significant deviation from 1, implies strong dependencies between data which do not change much with the string length (abscissa axis).

The approximated tau weights from Figure 4, though computed from the exact training information, consider only 2-data dependency, hence do not reflect behavior similar to those in Figure 3. Other calibration techniques are presently in development to mimic better the behavior of the exact tau weights.

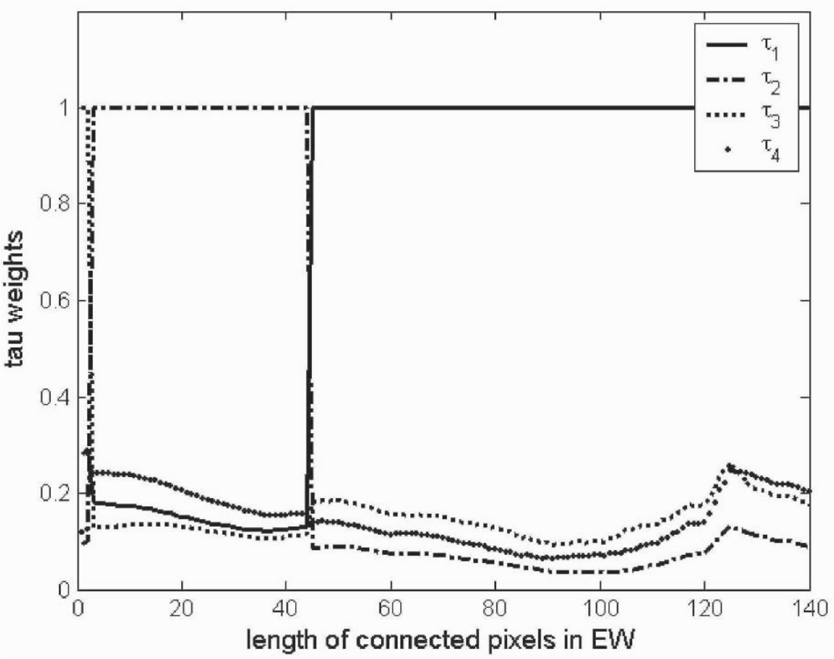


Figure 4 : Tau weights computed using the calibration technique

Impact of conditional independence

Conditional independence (CI) amounts to setting all tau weights equal to 1. We compare the estimate of P(A|D_1, D_2, D_3 ,D_4) resulting from CI with the true experimental probability obtained from the vector-Ti. Figure 5 shows this comparison. Note here that the tau estimate identifies the true probability since we have used an exact training information. Here, the assumption of CI ignores data redundancy, giving maximum importance to each datum. Therefore it results in an incorrect maximum compounding of information P(A|D_1, D_2, D_3 ,D_4) ≈ 1 at all lags: information are compounded so much that the combined probability reaches the maximum of 1 at all lags.

Many data combination algorithms for example Bayes nets (Pearl, 2000) and Markov chain-Monte Carlo based algorithms make extensive use of conditional independence. The maximum data compounding induced by such hypothesis may result in severe bias.

This page intentionally blank

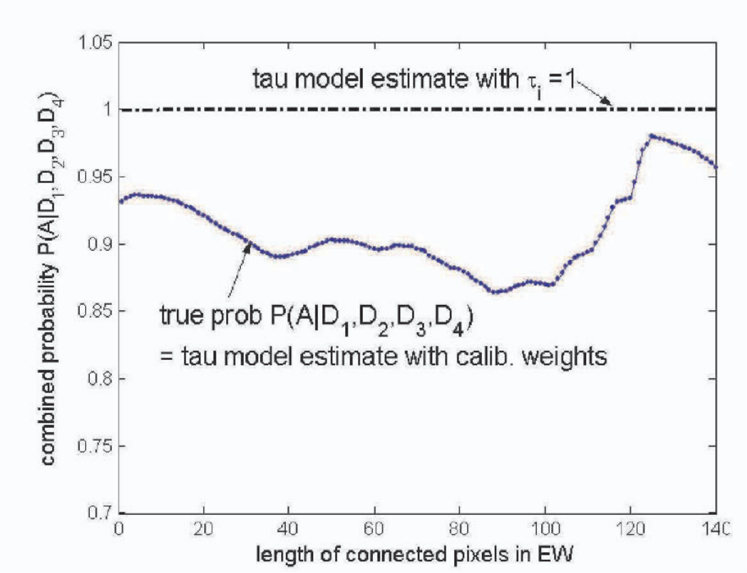


Figure 5: Comparing estimate from conditional independence (τ_i = 1) with tau estimate and the true probability $P(A|D_1,D_2,D_3,D_4)$

Conclusions:

This paper shows the generality of the tau model in handling complex dependencies between data. Here, the dependency between multiple-point data coming from different supports is handled using the tau weights. In general, this model can be used to combine probability information derived from any source, spatial or not. One can to devise innovative methods to represent the data redundancy for different examples. The tau model framework allows to account for joint dependency between data as opposed to 2-point or 2-data dependence typical of many estimation techniques. Therefore, efforts must be made to evaluate such joint data dependence. One way to do that is outlined in this paper through the concept of a vectorial training image. As shown here, ignoring this data dependence can be costly and making assumptions such as conditional independence are seldom safe.

References:

Bordley, R., 1982, *A multiplicative formula for aggregating probability assessments*, Management Science, V. 28, No. 10, pp. 1137-1148.

Journel, A.G., 2002, *Combining knowledge from diverse sources: An alternative to traditional data independence hypotheses*, Mathematical Geology, Vol. 34, No. 5, 573-596

Deutsch., C. V. and Journel, .A. G. , 1995, *GSLIB: A geostatistical software library and user's guide*, Oxford University Press.

Krishnan, S., 2004, Combining diverse and partially redundant information in the Earth sciences, PhD thesis, Stanford University.

Krishnan, S., Boucher. A. and Journel. A.G., 2004, *Evaluating information redundancy through the tau model*, Proceedings of 2004 Geostatistics Congress, Banff, Canada.

Krishnan, S., and A. G. Journel, *Spatial connectivity: from variograms to multiple-point measures,* Mathematical Geology, v. 35, Nov. 2003, pp. 915-925

Pearl, J., 2000, Causality: models, reasoning and inference, Cambridge University Press.

VALIDATION OF FIRST-ORDER STOCHASTIC THEORIES ON REACTIVE SOLUTE TRANSPORT IN HIGHLY STRATIFIED AQUIFERS

DANIEL FERNÀNDEZ-GARCIA and J. JAIME GÓMEZ-HERNÁNDEZ
Departamento de Ingeniería Hidráulica y Medio Ambiente,
Universidad Politécnica de Valencia, E.T.S.I. Caminos, Canales y Puertos, camino de vera s/n, 22012-46071 Valencia, Spain.

Abstract. Spatial variability in the physical and chemical properties of aquifers plays an important role in field-scale sorbing solute transport. Stochastic simulation of random functions can provide equiprobable maps of important aquifer parameters from a limited data set that can be use as input for making predictions of the movement and mixing of contaminated plumes. However, although this is the most powerful tool, this method requires large computational CPU-times. Stochastic solutions of flow and solute transport not only can be used to determine first-order approximations of the solution of the forward flow and transport problems but also can be conveniently employed to solve the inverse problem. However, these analytical solutions inherit very restrictive assumptions that need to be validated. Stochastic simulations of solute transport in highly stratified aquifers with spatially varying hydraulic conductivity and retardation factors were conducted to examine the validity of first-order stochastic analytical solutions for the travel time variance of breakthrough curves (BTCs) obtained at several control planes perpendicular to the mean flow direction. First, it is shown how to accurately calculate the temporal moments of BTCs and its ensemble average in particle tracking transport codes without having to evaluate the actual BTC from the distribution of particle travel times. Then, this methodology is used to evaluate how accurate small perturbation stochastic analytical solutions are. It is seen that theoretical stochastic predictions are valid for $\sigma^2_{\ln K}$ up to 1.0. For very heterogeneous aquifers, stochastic predictions will largely underestimate the travel time variance of BTCs.

1 Introduction

Solute transport in aquifers is greatly influence by natural heterogeneity. The concept of random functions offers a convenient way of describing the spatial variability in aquifer properties. For instance, stochastic simulation of random functions can provide equiprobable maps of hydraulic conductivity and retardation factors from a limited data set that can be used for making predictions of the movement and mixing of contaminated plumes. That is, these maps of aquifer properties can be used as input parameters in flow and solute transport solvers to generate equiprobable solutions of the transport problem. However, although this is the most powerful tool, this method requires large computational CPU-times. Under some restrictive assumptions, including stationary of the log-hydraulic conductivity (lnK) field, simple boundary conditions (infinite aquifer extension), negligible local dispersivity, uniform flow and mild

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 1117–1122.

heterogeneity (variance of lnK, $\sigma^2_{\text{lnK}} < 1$), analytical stochastic solutions of flow and solute transport are available. Despite the fact that these analytical solutions have enormously improved the knowledge of solute transport in heterogeneous aquifers, these analytical expressions not only can be used to determine first-order approximations of the solution of the forward problem but also can be conveniently employed to solve the inverse problem. In addition, they can also be used to verify flow and solute transport codes in heterogeneous aquifers.

Within this context, we examine the robustness of these analytical expressions by means of three-dimensional stochastic simulations of linearly sorbing reactive solute transport in highly stratified physically and chemically heterogeneous aquifer under different degrees of heterogeneity. Specifically, we compare the travel time variance of breakthrough curves (BTCs) obtained at control planes with stochastic first-order analytical expressions for highly stratified aquifers. It is shown that the simulated travel time variance is in good agreement with analytical expressions for σ^2_{lnK} up to 1.0.

2 Design of simulations

2.1 MATHEMATICAL STATEMENT

In steady-state flow conditions, linearly sorbing solute transport through heterogeneous porous media is governed by the following differential equation [*Freeze and Cherry*, 1979],

$$\phi R(\bar{x})\frac{\partial C}{\partial t} = -\sum_{i=1}^{3} q_i(\bar{x})\frac{\partial C}{\partial x_i} + \sum_{i=1}^{3}\sum_{j=1}^{3}\frac{\partial}{\partial x_i}\left(\phi D_{ij}\frac{\partial C}{\partial x_j}\right) \tag{1}$$

Where C is the dissolved concentration of solute in the groundwater, ϕ is the porosity, q_i is the ith component of Darcy's velocity, $q_i(\mathbf{x})=-K(\mathbf{x})\partial h/\partial x_i$, D_{ij} is the dispersion tensor, R is the retardation factor, h is the hydraulic head, and K is the hydraulic conductivity. Solute transport was simulated according to (1). Conservative tracers were only transported by advection and dispersion, whereas reactive tracers were further subjected to sorption processes represented by a reversible linear equilibrium isotherm which considers that sorbed solute mass is proportional to the concentration of dissolve solute by a factor known as the distribution coefficient K_d. This is frequently the case for many non-polar organic hydrophobic substances dissolved in groundwater. The retardation factor is expressed as $R(\mathbf{x})=1+K_d(\mathbf{x})\rho_b/\phi$. Where ρ_b is the bulk density of the soil.

The natural logs of hydraulic conductivity lnK($\mathbf{x}$) and distribution coefficient $\ln K_d(\mathbf{x})$ were two spatially correlated field variables. The lnK($\mathbf{x}$) variable was perfectly correlated with the $\ln K_d(\mathbf{x})$ variable assuming a linear negative model. Field observations [*Garabedian et al.*, 1988] as well as laboratory column and batch experiments [*Fernàndez-Garcia et al.*, 2004a] suggest that sorption properties in an aquifer are spatially variable. For instance, distribution coefficients for strontium on 1,279 subsamples of cores from the Borden aquifer gave K_d values that ranged from 4.4 ml/g to 29.8 ml/g with a geometric mean of K_d 0.526 ml/g and standard deviation of the

$\ln K_d$ of 0.267 [*Robin et al.*, 1991]. A statistically significant (at the 99.95%) but very weak negative correlation between the $\ln K_d$ and lnK was observed.

2.2 NUMERICAL FEATURES

The computational domain is discretized into a regular mesh formed by 250×250×200 parallelepiped cells. The heterogeneous structure of the hydraulic conductivity field resembles those from the Borden aquifer [*Mackay et al.*, 1986; *Sudicky*, 1986]. The $\ln K(\mathbf{x})$ is a second-order stationary multi-Gaussian random field with anisotropic exponential covariance function defined with horizontal correlation scale $\lambda_H = 2.78$ m and vertical correlation scale $\lambda_V = 0.278$ m. The computational resolution is of five grid cells within a correlation scale in all directions. Retardation factors were estimated as $R(\mathbf{x})=1+\rho_b/\phi K_{dg}\exp[af(\mathbf{x})]$, with $f(\mathbf{x})$ being the fluctuation of $\ln K(\mathbf{x})$ around the mean. K_{dg} is the geometric mean of K_d, and a reflects the relationship between $\ln K_d(\mathbf{x})$ and $\ln K(\mathbf{x})$. Here, K_{dg}, ρ_b, and ϕ were similar to the Borden aquifer with values of 0.526 ml/g, 1.81 g/cm^3, and 0.35, respectively. The parameter a was set to –0.5. When sorption processes are completely linked to grain surface areas, *Garabedian et al.* [1988] showed that a power law relationship between conductivity and mean grain radius yields a=–0.5.

The hydraulic conductivity field is incorporated into a seven-point finite difference ground-water flow model, MODFLOW2000 [*Harbaugh et al.*, 2000]. Upstream and downstream boundaries are specified as constant heads, such that the hydraulic gradient in the mean flow direction is 0.004. No-flow conditions are prescribed at the transverse, top and bottom boundaries. The model calculates the flow rates at the grid interfaces, which yields the velocity field. Porosity is spatially homogeneous with a value of 0.35. This velocity field is then used in a Random Walk Particle Tracking transport code similar to the one by *Tompson* [1993] and *Wen and Gómez-Hernández* [1996] that simulates the solute migration by partitioning the solute mass into a large number of representative particles (the number of particles used for all simulations is 10 000); moving particles with the velocity field simulates advection, whereas a Brownian motion is responsible for dispersion. Local longitudinal dispersivity was set to 2.78 cm and the ratio of the longitudinal to the local transverse dispersivity is 1/10. Molecular diffusion was neglected. Initially, the particles are randomly distributed in a plane transverse to the mean flow direction. This plane is located three correlation scales away from the upgradient boundary to avoid boundary effects. The shape of the particle source is a rectangle centered within this plane. The source size is of 40-correlation scales in the transverse direction to the mean flow and 30-correlation scales in the vertical direction. This leaves a gap of 5-correlation scales between boundaries and source.

2.3 EVALUATION OF TRAVEL TIME VARIANCE

Monitoring the first passage time of particles passing through control planes allows for the estimation of BTC temporal moments without having to evaluate the actual shape of the BTC. The nth-absolute temporal moment can be calculated as the expected value of the arrival time of a particle at the control plane to the nth power [*Fernàndez-Garcia,* 2003; *Fernàndez-Garcia et al.*, 2004b],

$$\mu_n^{'}(x_1) = \frac{1}{m}\int_0^{\infty} t^n Q C_f(x_1,t)dt \approx \frac{1}{NP_a}\sum_{k=1}^{NP_a}\left(t_p^{(k)}(x_1)\right)^n \tag{2}$$

Where Q is the total water flux passing through the control plane, m is the total mass injected, x_1 is the mean flow direction coordinate, C_f is the flux concentration of solute passing through a given surface, $t_p^{(k)}$ is the first arrival passage time of the kth particle, and NP_a is the total number of particles arrived at the x_1-control plane. The advantage of this methodology is that it avoids constructing the entire BTCs and subsequent integration of concentrations over time in (2). Evaluation of the entire BTC from the distribution of particle arrival times at control planes requires smoothing techniques which would originate larger numerical errors. The nth-absolute temporal moment of the ensemble average BTCs is simply the average of the nth-absolute temporal moments across all realizations [*Fernàndez-Garcia et al.*, 2004b],

$$M_{T,n}^{'}(x_1) = \frac{1}{m}\int_0^{\infty} t^n \langle QC_f(x_1,t)\rangle dt = \langle \mu_n^{'}(x_1)\rangle \approx \frac{1}{N_r}\sum_{m=1}^{N_r}\mu_n^{'(m)}(x_1) \tag{3}$$

Where $\mu^{'}_n(x_1)$ is the nth-absolute temporal moment obtained at the x_1-control plane in a single realization of the aquifer, N_r is the number of realizations (20), the brackets <·> denotes the ensemble average, and the superscript m indicates the realization number. The mean arrival time $T_a(x_1)$ and the travel time variance $\sigma^2_t(x_1)$ of the ensemble average BTC is calculated as

$$T_a(x_1) = M_{T,1}^{'}(x_1) \qquad \sigma_t^2(x_1) = M_{T,2}^{'}(x_1) - \left(M_{T,1}^{'}(x_1)\right)^2 \tag{4}$$

3 Simulations results

For highly stratified porous media with large anisotropy ratios in the correlation scale of the heterogeneous structure ($\lambda_H/\lambda_V > 10$) and for small $\sigma^2_{\ln K}$ and σ^2_R, the ratio of the variance of the particle position to the mean travel distance can be written as [*Dagan*, 1989; *Dagan and Cvetkovic*, 1993; *Miralles-Wilhelm and Gelhar*, 1996]

$$A(\xi) = \frac{\sigma_x^2(\xi)}{2\xi} = \sigma^2_{\ln K}\lambda_H^2\left[1 + \frac{(\exp(-\xi b(e)) - 1)}{\xi b(e)}\right]\left[1 - \frac{a}{R_A}\left(\frac{\rho_b}{\phi}K_{dg}\right)\right]^2 \tag{5a}$$

$$b(e) = 1 + \frac{19e^2 - 10e^4}{16(e^2-1)^2} - \frac{e(13-4e^2)\arcsin(\sqrt{1-e^2})}{16\sqrt{1-e^2}(e^2-1)^2} \tag{5b}$$

Where R_A is the mean retardation factor, e is the ratio of vertical to the horizontal correlation scale, ξ is the mean travel distance of the center of mass normalized by the horizontal correlation scale, and a is the parameter that correlates the $\ln K(\mathbf{x})$ and $\ln K_d(\mathbf{x})$

random fields as given in section 1.2. In essence, the parameter A(ξ) is an operational dispersivity value that can be used to replace the heterogeneous aquifer with an equivalent homogeneous porous media. Operational dispersivity values obtained from stochastic simulations used for comparison with (5) were derived from the mean arrival time and travel time variances of ensemble average BTCs as

$$A(\xi)=\frac{\xi}{2}\frac{\sigma_t^2(\xi)}{\left(T_a(\xi)\right)^2} \qquad (6)$$

Where ξ is the distance from the particle source to the x_1-control plane normalized by the horizontal correlation scale. It should be noted that operational dispersivity values estimated from particle travel times using (6) and those estimated from particle spatial location using (5) are equivalent for small $\sigma^2_{\ln K}$ and σ^2_R [*Fernàndez-Garcia et al.*, 2004b].

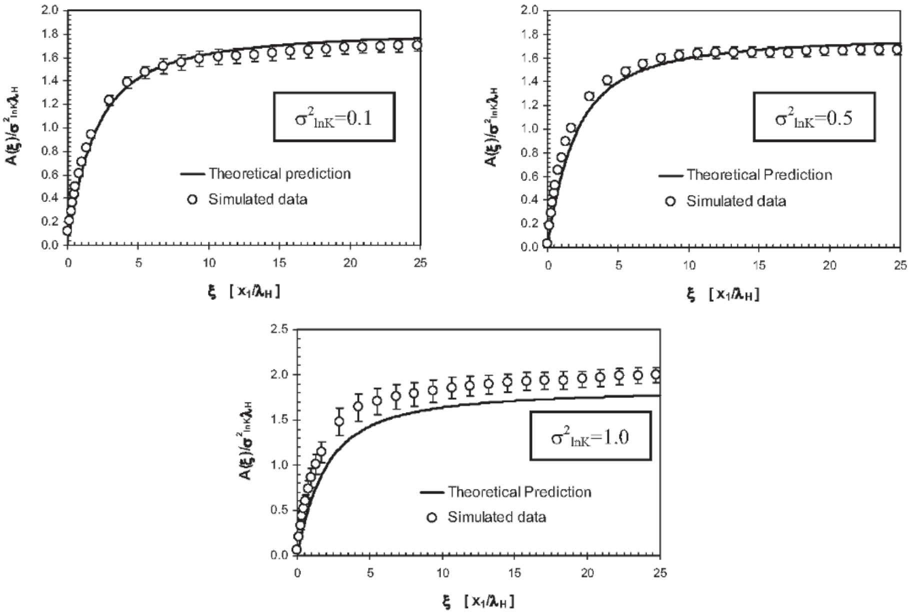


Figure 1. Comparison of simulated operational dispersivity values obtained from particle travel times at x_1-control planes with small perturbation theoretical predictions. Error bars represent the 95% confidence interval of the mean simulated value.

Figure 1 compares simulated operational dispersivity values A(ξ) of the ensemble average BTCs obtained at different x_1-control planes with analytical stochastic solutions for highly stratified aquifers. It is seen that stochastic simulations of solute transport are in perfect agreement with analytical solutions for $\sigma^2_{\ln K}$ up to 0.5 and they are in reasonably good agreement for $\sigma^2_{\ln K}$ up to 1.0 where simulated A(ξ) values start deviating from its theoretical prediction. In addition, for very heterogeneous aquifers

($\sigma^2_{\ln K}$ > 1.0), it is shown that analytical expressions will underestimate the actual operational dispersivity value obtained from BTCs.

4 Conclusions

Stochastic simulations of solute transport in highly stratified aquifers with spatially varying hydraulic conductivity and retardation factor were conducted to examine the validity of first-order stochastic analytical solutions for the travel time variance of BTCs obtained at several control planes perpendicular to the mean flow direction. First, it is shown how to accurately calculate the temporal moments of BTCs and its ensemble average in particle tracking codes without having to evaluate the actual BTC from the distribution of particle travel times in heterogeneous aquifers. Then, this methodology is used to evaluate how accurate can be the small perturbation stochastic analytical solutions. It is seen that theoretical stochastic predictions are valid for $\sigma^2_{\ln K}$ up to 1.0, but for very heterogeneous aquifers, stochastic predictions will largely underestimate the travel time variance of BTCs.

References

Dagan, G. *Flow and Transport in Porous Formations*, Springer-Verlag, 465 p., 1989.

Dagan, G., and V. Cvetkovic, 1993. Spatial moments of a kinetically sorbing solute plume in a heterogeneous aquifer. *Water Res. Res.*, 29, 4053-4061.

Fernàndez-Garcia. *Scale-dependence of Non-reactive and Sorptive Transport Parameters Estimated from Radial and Uniform Flow Tracer Tests in Heterogeneous Formations: Experimental and Numerical Investigations*. Ph.D. Thesis. Colorado School of Mines, 396 p., 2003.

Fernàndez-Garcia, D., T. H. Illangasekare, and Harihar Rajaram, 2004a. Conservative and sorptive forced-gradient and uniform flow tracer tests in a three-dimensional laboratory test aquifer. *Water Res. Res*., 40, W10103, doi:10.1029/2004WR003112.

Fernàndez-Garcia, D., T. H. Illangasekare, and Harihar Rajaram, 2004b. Differences in the scale dependence of dispersivity estimated from temporal and spatial moments in chemically and physically heterogeneous porous media. *Advances in Water Resources*, Accepted.

Freeze, R. A., and J. A. Cherry, *Groundwater*, Prentice-Hall, Englewood Cliffs, N. J., 1979.

Garabedian, S. P., Gelhar, L. W., and Celia, M. A., Large-scale dispersive transport in aquifers: Field experiments and reactive transport theory, *Rep.* 315, Ralph M. Parsons Lab., Dep. of Civil Eng., Mass. Inst. of Technol., Cambridge, 1988.

Harbaugh, A. W., Banta, E. R., Hill, M. C., and McDonald, M. G. MODFLOW-2000, *The U.S. Geological Survey Modular Ground-Water Model–user guide to modularization concepts and the ground-water flow process*, Open-file Report 00-92, 2000.

Mackay, D. M., Freyberg, D. L., Roberts, P. V., Cherry, J. A., 1986. A Natural Gradient Experiment on Solute Transport in a Sand Aquifer, 1. Approach and Overview of Plume Movement. *Water Res. Res.*, 22 (13), 2017-2029.

Miralles-Wilhelm, F., L. W. Gelhar, 1993. Stochastic analysis of sorption macrokinetics in heterogeneous aquifers. *Water Res. Res.*, 32(6), 1541-1549.

Robin, M. J. L., Sudicky, E. A., Gillham, R. W., Kachanoski, R. G., 1991. Spatial variability of strontium distribution coefficients and their correlation with hydraulic conductivity in the Canadian forces base Borden aquifer. *Water Res. Res.*, 27 (10), 2619-2632.

Sudicky, E. A, 1986. A natural gradient experiment on solute transport in a sand aquifer: spatial variability of hydraulic conductivity and its role in the dispersion process. *Water Res. Res.*, 22(13), 2069-2082.

Tompson, A. F. B., 1993. Numerical simulation of chemical migration in physically and chemically heterogeneous porous media. *Water Res. Res*., 29(11), 3709-3726.

Wen, X.-H., J. J. Gómez-Hernández, 1996. The constant displacement scheme for tracking particles in heterogeneous aquifers. *Ground Water*, 34(1), 135-142.

GEOSTATISTICAL AND FOURIER ANALYSIS APPROACHES IN MAPPING AN ARCHAEOLOGICAL SITE

ABÍLIO A.T. CAVALHEIRO and JORGE M.C.M. CARVALHO
Department of Mining & Geoenvironmental Engineering - CIGAR
Faculty of Engineering - University of Porto
R. Dr. Roberto Frias, 4200-465 Porto - Portugal

Abstract. The role of geophysical methods for assessing archaeological sites has been increasing in recent years. One of the most commonly applied is the electrical resistivity method, responding to differences in underground electrical conductivity. The raw field data resulting from such surveys are often difficult to interpret due to the combined effects of regional low frequency trends and/or high frequency noise. The traditional treatment of signal based on Fourier analysis, namely the use of filtering techniques is sometimes used to improve the quality of the signal. Assuming the bandlimited condition of the sampled function, a bandlimited interpolator can be the option as estimation method. Alternatively or in parallel with the mentioned approach, geostatistical tools may be a useful methodology in modelling the spatial variability of the regionalised geophysical variable as well as in interpolation leading to the final site mapping. In this study the referred different approaches are applied to data resulting from a geophysical survey using the resistivity method and a topographic survey in two areas located in the northern part of Portugal in which, most probably, archaeological structures are buried. Some of the obtained results using inverse-square distance, kriging and band-limited interpolation, are herein presented and compared.

1 Introduction

Old constructions of our human ancestors, namely for burial purposes, were often progressively covered by soil and, in present days, most of them are hidden underground although sometimes suspected through a typical topographic signature ("anomalous" elevation). Geophysical methods are sometimes applied in order to investigate places where archaeologists found evidences of possibly existing structures, aiming at excluding from excavation most unlikely site areas or with the purpose of obtaining an estimation of the extension of the archaeological site for planning future work.

The often existing contrast in electrical conductivity/resistivity between the buried manmade stone structures and the surrounding usually more conductive soil, point to the adequacy of electrical methods to detecting such buried structures.

Various geological parameters influence the mentioned contrast in resistivity, namely porosity, water content and degree of saturation. Lower porosity stones, will tend to have lower water content and consequently higher resistivity.

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 1123–1129.

In this context, one of the most used electrical methods is the so-called resistivity method used to determine from surface readings the underground resistivity distribution. Basically, two current electrodes, A and B, are used to inject electrical current into the ground and two potential electrodes, C and D, are used for measuring the resulting potential difference (Figure 1). The intensity of current and potential readings, together with the geometric electrode configuration, allow calculating the so-called apparent resistivity value, meaning the resistivity of a homogeneous ground in which would be obtained identical readings as those obtained in the studied inhomogeneous terrain with the same geometric electrode configuration. The array AMNB may be successively moved along a traverse with each new reading being assigned to the central point of the array. Several parallel traverses allow defining a 2D surface of resistivity readings assumed situated at a certain depth function of the array length. In Figure 1 is sketched the so-called Wenner-alpha array of equally spaced electrodes, a buried structure and the correspondent resistivity anomaly.

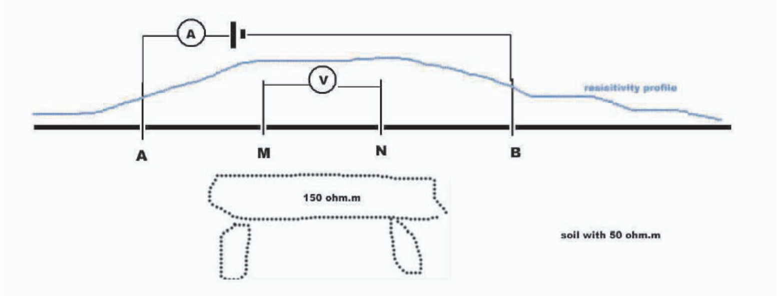


Figure 1. The Wenner-alpha array and a resistivity profile over a buried archaeological structure.

The detection aptitude of the method, including signal to noise ratio, depends on several factors namely the conductivity contrast between the structure and the surrounding soil, the burial depth and size of the structure and the electrode configuration and sampling spatial rate.

There are different electrode arrays and field procedures that may be used to obtain 2D or 3D images of the underground, combining profiling with vertical electrical soundings, leading to apparent resistivity vertical sections. Then, data treatment based on inversion algorithms often creates very realistic models of the hidden underground reality. Nevertheless, in many practical situations the only used information is the one obtained in a regular surface grid survey leading to 2D horizontal mapping.

The improvement of 2D resistivity horizontal mapping, using Kriging, Fourier analysis and filtering techniques is the main aim of this paper.

2 Notes on interpolation

The common situation when interpolation is used results from the need to reconstruct some time and/or space signal based on discrete readings/samples. Interpolation possibilities increased a lot with the progressive spread of digital equipment. Analogous devices interpolation was physically done, for example the needle tracing a graph over a paper, the light impression of the salt silver grains of the radiographic plate and so on. When such devices were substituted by digital equipment the signals, including images, were reconstructed using sampled discrete information. Nowadays a very common reconstruction device is the cellular phone, using numerically digitized samples of the human voice. So we may say that the digitized devices increased the use of interpolation techniques and concomitantly a variety of algorithms were developed in different areas. Given the pervasive characteristic of the problem, those techniques are spreading to a number of different disciplines, as is the case of geostatistical kriging methods.

3 Fourier analysis point of view versus Kriging

If the readings are made in the field at a regular sampling step, let's say Δl meters, and consequently Fourier analysis may be easily used, Shannon's sampling theorem states that the maximum frequency (Nyquist frequency) present in the obtained signal is 1/(2 Δl) cycles/m, so a proper interpolation should not introduce higher frequencies than those present in the original sampled signal, taking in consideration that those higher frequencies would only be present if the sample step had been smaller when acquiring the original data in the field.

Obviously when in practice interpolation takes place a combination of accurately reconstructed and "invented" information is produced but, as much as possible, the used procedures shall be based on some kind of optimized sense. In the Fourier analysis perspective the interpolation process must not introduce higher frequencies (bandlimited interpolation) than those maximum frequencies allowed by the sampling theorem (Papoulis, A., 1962). Otherwise the interpolated data will undergo some kind of aliasing when compared with the data that would be obtained with a higher sampling rate. On the other hand, the smoothing generic effect (high frequencies cutting) of several interpolation procedures namely kriging, is also an inconvenient feature. Kriging interpolation uses the auto-correlogram that may be obtained from the Fourier power spectrum, loosing the phase information (Bracewell, R., 1978). Apparently kriging does not need to use an original equally spaced grid data. Being more exact, in fact kriging results based on semi-variogram values, allow a kind of tolerance (Goovaerts, P., 1997) enabling the use of a not equally spaced grid of raw-data. Even though, semi-variograms calculated values benefits of higher accuracy if the original data is equally-spaced. Both methods (bandlimited and kriging interpolation) implicitly accept the additive hypothesis - linear superposition of the measures of the studied property – since, in both cases, to obtain an interpolated point value a linear combination of original reading values is used. One may state that kriging and bandlimited interpolation have common characteristics in their mathematical foundations. However, the authors believe that Fourier analysis allows an enriching insight into geostatistical estimation.

4 Megalithic dolmens ("mamoas")

An often used indication of the presence of archaeological vestiges of megalithic monuments is the "anomalous" elevation of the topographic surface appearing in an even surface ground. In Portugal such features are called "mamoas". Under the resistivity perspective a mamoa is a combination of less conductive buried stones having a non fortuitous shape and geometry, surrounded by more conductive covering soil. The final result is a high resistivity anomaly that may be detected by resistivity surveying. Two mamoas situated on the northern part of Portugal were studied: mamoa-A, located in Vale da Porca – Trás-os-Montes, and mamoa-B at Castro Laboreiro - Minho. The resistivity survey was carried out with an ABEM SAS 300C Terrameter resistivimeter.

To graphically represent the resistivity and topographic surfaces, Surfer © software shaded relief images overlaid by contour maps were used. The graphical data representation, without interpolation, that is, using a grid with only the original data points are shown in Figures 2 and 3. For Mamoa-A the rectangular grid opposite corners are (2.75, 1.5), (17.75, 25.5), sampling step 1.5 m, and for Mamoa-B (0,0), (20,22), sampling step 1 m.

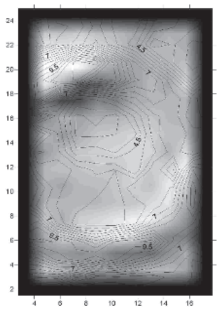


Figure 2. Mamoa-A: original electrical resistivity data image.

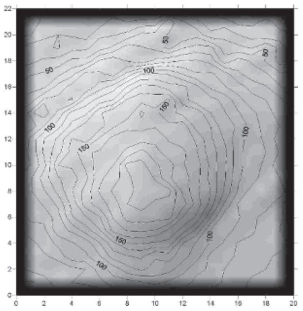


Figure 3. Mamoa-B: original topographic data image.

Previously to kriging interpolation, a structural analysis was carried out using the software Variowin (Pannatier Y., 1996). In Figures 4 and 5 are respectively the variogram surface and omnidirectional variogram model fit corresponding to mamoa-A.

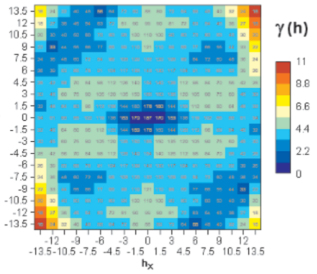


Figure 4. Mamoa-A: variogram surface.

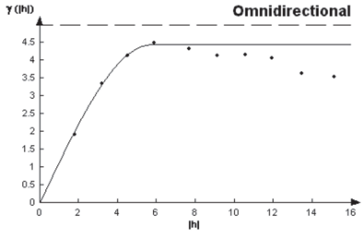


Figure 5. Mamoa-A: omidirectional spherical variogram fit.

The correspondent graphs for mamoa-B can be seen in Figures 6 and 7, having the variogram surface a distinctive pattern. In case of mamoa-A the shown graphs are related to field data prior to being transformed in apparent resistivity values through multiplication by the constant Wenner-alpha geometric factor, K = 2πa (a = constant distance between electrodes).

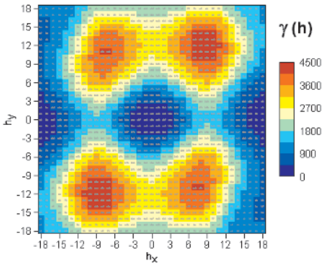


Figure 6. Mamoa-B: variogram surface.

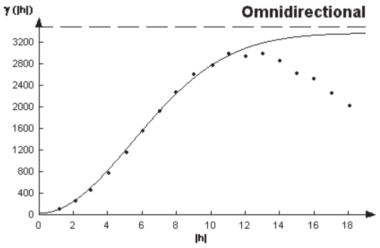


Figure 7. Mamoa-B: omidirectional gaussian variogram model fit.

Next Figures, 8 to 13, show the results of different interpolation methods used to increase 10 times the correspondent original grids density, so that for mamoa-A the x and y ticks became 0.15 m and for mamoa-B 0.1m. The very spiky aspect of the images obtained with the inverse square distance method (ISD), that may be seen in Figures 8 and 11, is an evident sign of the poor performance of ISD method. Figures 9 and 12 were obtained with a 2D lowpass bandlimited algorithm (LPBL) adapted in Matlab environment using the Signal Processing Toolbox "interp.m" function. Figures 10 and 13 were obtained using ordinary kriging (OK) with the respective omnidirectional variogram models. Note that all the used methods of interpolation reproduce the original values whenever any interpolated grid node matches a location of an original data value.

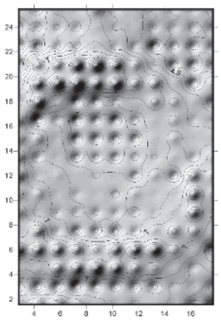


Figure 8. Mamoa-A: ISD map.

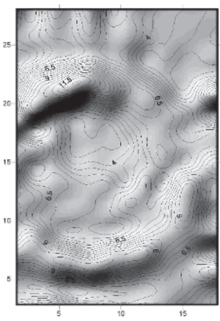


Figure 9. Mamoa-A: LPBL map.

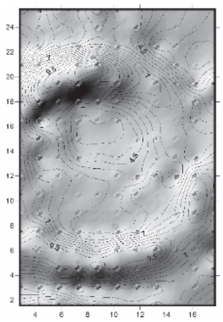


Figure 10. Mamoa-A: OK map.

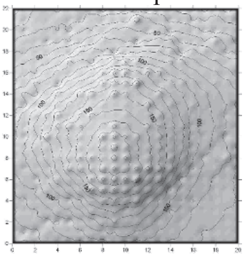


Figure 11. Mamoa-B: ISD map.

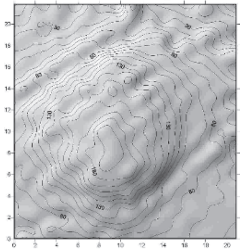


Figure 12. Mamoa-B: LPBL map.

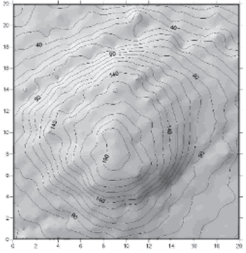


Figure 13. Mamoa-B: OK map.

Kriging and bandlimited interpolation may be compared using a 2D Fourier transform. As an example for mamoa-A original data, Figures 14 and 15 were obtained using a 2D fast Fourier transform, FFT, and both show all the transform module graphs, respectively along the x and y directions, against the amplitude axis. Along x direction, two graphs of the module of the transforms of interpolated data were obtained, as shown in Figures 16 and 17.

All the Fourier transforms of Figures 14 to 17 were obtained after subtracting from each value the data global mean value. It may be seen comparing Figures 16 and Figure17, that lowpass bandlimited and kriging interpolation produced very similar spectra, yet having this last one some more high frequency noise content, as it could be expected comparing visually the maps from Figures 9 and 10.

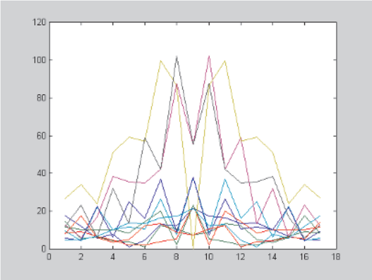


Figure 14. Mamoa-A original data: x direction spectra.

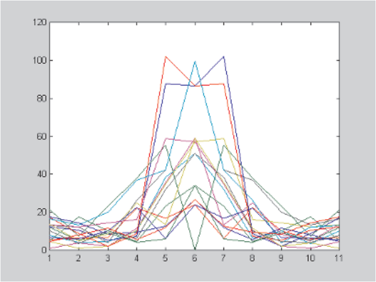


Figure 15. Mamoa-A original data: y direction spectra.

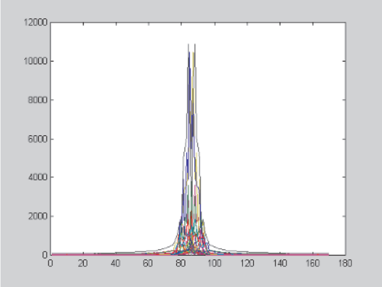


Figure 16. Mamoa-A LPBL interpolation: x direction spectra.

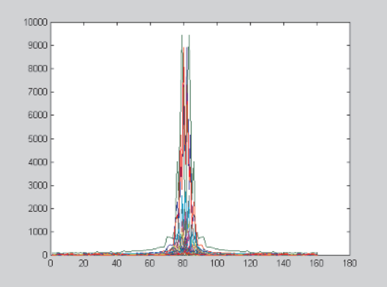


Figure 17. Mamoa-A kriging interpolation: x direction spectra.

5 Conclusions

Bandlimited interpolation has the advantage of generating denser grids maintaining the original frequency content of the original data. Fourier analysis identifies, as well as the variogram representation, the structure of spatial correlation. Its main inconvenience is the need of data assuming regularly spaced readings, a hindrance that apparently kriging practically overcomes.

The inverse square distance method maps were of poor quality.

This page intentionally blank

Kriging introduced some higher noise in higher frequencies when compared with the bandlimited interpolator. The Fourier analysis may be a very useful conceptual tool to help understanding the advantages and limitations of geostatistical procedures.

References

Bracewell, R., *The Fourier Transform and its Application*, Mc Graw-Hill, 1978.
Papoulis, A., *The Fourier Integral and its Applications*, Mc Graw-Hill, 1962.
Goovaerts, P., *Geostatistics for Natural Resources Evaluation*, Oxford University Press, 1997.
Journel, A.G. and Huijbregts, C.J., *Mining Geostatistics*, Academic Press, 1978.
Pannatier, Y., *Variowin – Software for Spatial data Analysis*, Springer, 1996.
Krauss, Thomas P., Loren Shure and John N. Little, *Signal Processing Toolbox User's Guide*,
The Mathworks, Inc. 1994

BATGAM$^{©}$ GEOSTATISTICAL SOFTWARE BASED ON GSLIB

BRUCE BUXTON, ALAN PATE, MICHELE MORARA
Battelle Memorial Institute
Measurement and Data Analysis Sciences
505 King Avenue
Columbus, Ohio 43201

Abstract. GSLIB is a relatively inexpensive (it is free!) yet flexible and useful software library for performing many geostatistical analyses, such as semivariogram calculations, kriging, and conditional simulation (Deutsch and Journel, 1998, Oxford University Press). Unfortunately, GSLIB provides only FORTRAN source code and executables, and the user is required to work with text parameter files or develop his/her own interface to input data and output results. To help alleviate this situation, Battelle has developed the BATGAM$^{©}$ software which provides just such an interface in a user-friendly Windows environment.

Battelle has not attempted to implement all of the routines included in GSLIB, but has first focused on the most commonly required modules for typical site characterization studies, namely semivariogram calculation and modeling, kriging, and conditional simulation. The first program implemented was GAMV for three-dimensional semivariogram calculations based on irregularly-spaced (i.e., non-gridded) data. BATGAM$^{©}$ provides an interface for inputting the user's data and generating an output file of semivariogram values. BATGAM$^{©}$ also provides a flexible modeling module, based on GSLIB's COVA subroutine, to fit and graphically display the semivariograms with their models.

The second program implemented in BATGAM$^{©}$ was KT3D for three-dimensional kriging of points or blocks using a variety of kriging options. In addition to interfacing with the user's input data and calculating a gridded file of kriged estimates, BATGAM$^{©}$ also includes a simple module for color display of the kriging results.

The third GSLIB program implemented in BATGAM$^{©}$ was SGSIM for three-dimensional, sequential Gaussian conditional simulation of continuous variables. In addition, three other auxiliary programs have been implemented with SGSIM because they are useful when conducting conditional simulations. These programs are NSCORE, for performing a normal-score transformation with the original input data; BACKTR, for back-transforming simulated data via the inverse relationship associated with NSCORE; and GAM, for calculating semivariogram values based on large data sets with gridded data, such as one might construct in a conditional simulation. Similar

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 1131–1136.

to the kriging module, this simulation module generates gridded data and interfaces with a simple color graphics tool for displaying the simulated results.

1 Introduction

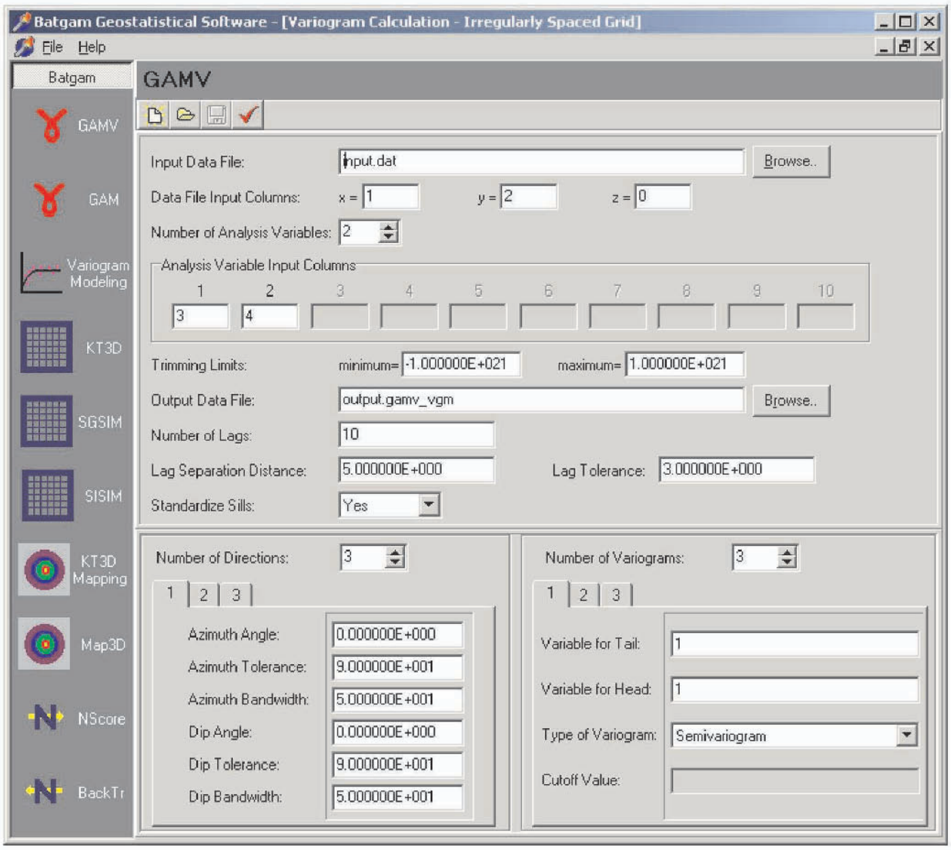
Over the past twenty years, Prof. André Journel, Prof. Clayton Deutsch, numerous Stanford University graduate students, and other colleagues have built a suite of geostatistical software routines to perform a wide variety of quantitative analyses. Many of these routines were released in 1992 (with revisions in 1998) as the GSLIB library of FORTRAN source code (Deutsch and Journel, 1998). This software is not meant to compete with commercially available computer mapping packages, but rather to provide a training tool and set of seed software with which individuals or companies could build their own geostatistical computing capabilities. However, since it is seed software, GSLIB does not generally include the important user interfaces needed to input data to the geostatistical algorithms and output and has limited options for displaying results for interpretation.

Battelle's BATGAM© software helps fill this void, at least for a few of the more commonly used geostatistical routines. Over the years we implemented various components of GSLIB, initially just with simple FORTRAN front ends and then more recently in a user-friendly Windows environment. BATGAM© is not flashy, but it is quite functional and remains faithful to the underlying GSLIB conventions. To help further the goals of Prof. Journel and make geostatistical software more widely available (thereby helping to promote the growing field of geostatistics), Battelle recently decided to make its BATGAM© software available free of charge. To obtain a complimentary copy, simply visit the environment link on our Web site at: **http://www.battelle.org** and/or contact one of the co-authors of this paper. Similar to GSLIB, Battelle does not support BATGAM© and provides no warranty as to its accuracy or usefulness. There is no BATGAM© user's guide; since we follow exactly the GSLIB conventions, we recommend purchasing a copy of the GSLIB book (Deutsch and Journel, 1998) to serve as a user's guide.

GSLIB contains a wide variety of geostatistical routines, ranging from simple utility programs and display routines, to semivariogram and kriging analyses, and a variety of sophisticated simulation algorithms. BATGAM© currently includes only those common routines needed to perform semivariogram, kriging, and conditional simulation analyses, as well as to view the results. Our plans are to gradually add other modules to BATGAM© when we find them useful for our environmental work.

2 BATGAM© Modules

BATGAM© currently includes the following modules which are accessed via a simple opening screen: GAMV (to calculate experimental semivariogram results), GAMV Modeling (to fit semivariogram models), KT3D (to calculate kriging estimates), KT3D Mapping (to display and map the kriging results), SGSIM (to perform a sequential Gaussian conditional simulation), GAM (to calculate semivariogram results with large, gridded data sets), NSCORE (to perform a normal-score transformation with original input data), and BACKTR (to back-transform simulated data via the relationship associated with NSCORE). By clicking the icon down the left side of the main BATGAM© screen that is associated with any given module, the user is guided to a tabular screen that requests input for each parameter associated with that module. For example, Figure 1 shows the parameter screen associated with the GAMV module. Each of the parameters corresponds to the same inputs that are listed in Deutsch and Journel (1998). For example, the parameters shown in Figure 1 correspond to the same inputs that are described on pages 53-55 of Deutsch and Journel (1998).



Input Data File: input.dat

Data File Input Columns: x = 1 y = 2 z = 0

Number of Analysis Variables: 2

Analysis Variable Input Columns

1	2	3	4	5	6	7	8	9	10
3	4								

Trimming Limits: minimum= -1.000000E+021 maximum= 1.000000E+021

Output Data File: output.gamv_vgm

Number of Lags: 10

Lag Separation Distance: 5.000000E+000 Lag Tolerance: 3.000000E+000

Standardize Sills: Yes

Number of Directions: 3

Azimuth Angle: 0.000000E+000

Azimuth Tolerance: 9.000000E+001

Azimuth Bandwidth: 5.000000E+001

Dip Angle: 0.000000E+000

Dip Tolerance: 9.000000E+001

Dip Bandwidth: 5.000000E+001

Number of Variograms: 3

Variable for Tail: 1

Variable for Head: 1

Type of Variogram: Semivariogram

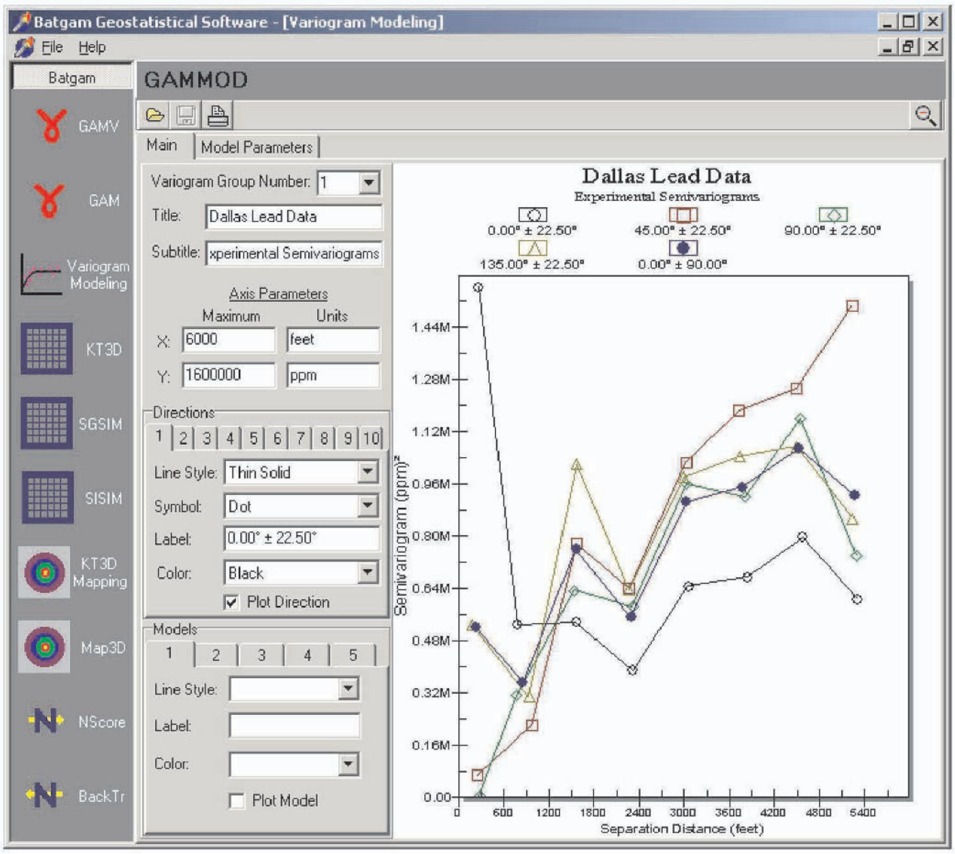
Cutoff Value:

Figure 1. GAMV Module for Experimental Semivariogram Calculations

GAMV is the module with which a user calculates experimental semivariogram values from his/her data set. As shown in Figure 1, the required user inputs are of two general types: those describing the user's data set and data coverage (top half of the screen), and those describing the specific kinds of experimental semivariogram calculations which are needed (bottom half of the screen).

The file handling conventions for each BATGAM© module follow simple Windows conventions, and are controlled by several buttons in the top-left corner of each BATGAM© screen. The 'File' button offers the user options to reset the screen (either by adopting a set of default parameter values, or by retrieving a previously saved set of parameter values), or to save the current parameter values and close out of the screen. Three other buttons (with blank-page, open-folder, and diskette icons) serve the same functions in a slightly different format. A fifth button, with a checkmark icon, instructs BATGAM© to run the requested module using the parameter values that are listed on the screen.

The file of experimental semivariogram calculations is accessed (via the Open Folder button) for graphical display and modeling through the **GAMV Modeling** screen. As shown in Figure 2, this step in the analysis is broken down into two parts (accessed via two tabs): graphical display on the Main screen, and semivariogram modeling on the Model Parameters screen.



Variogram Group Number: 1

Title: Dallas Lead Data

Subtitle: xperimental Semivariograms

Axis Parameters

	Maximum	Units
X:	6000	feet
Y:	1600000	ppm

Directions

1 2 3 4 5 6 7 8 9 10

Line Style: Thin Solid

Symbol: Dot

Label: 0.00° ± 22.50°

Color: Black

☑ Plot Direction

Models

1 2 3 4 5

Line Style:

Label:

Color:

☐ Plot Model

Figure 2. Main Screen of GAMV Modeling Module

The Model Parameters screen allows the user to specify up to five different semivariogram models, either simply to view on the Main screen, or more typically to overlay with a set of experimental semivariogram points for model fitting purposes. The graphical appearance of the models and the specification of whether to plot the models or not is determined by the user in the Main screen.

Kriging is performed in BATGAM© via the **KT3D** screen which has four associated tabs that request different kinds of input parameters defining the specific analysis needed by the user. The first tab (Main) specifies the user's file naming selections, various characteristics about the input data set, the kriging grid size and spacing, and the strategy for searching the data during kriging. The second tab (Variogram) requests the user's semivariogram model in the same format as it is described in the GAMV Modeling module. There is also a convention included at the top of the second tab to retrieve a file with semivariogram model parameters from an existing file that was written by GAMV Modeling. The third and fourth KT3D tabs (Drift and Jackknife) request specialized inputs when the user intends to perform kriging with a drift model or to jackknife the data, respectively.

After kriging, the user can visualize his/her results via the **KT3D Mapping** module. This is a simple program that generates color contour maps with the kriging output, either to be displayed on the screen, or to be outputted to a file for subsequent hardcopy printing. In addition to the color map, the user can request a simple contour line map. And in either case, the user has some control over the specific contour intervals that are utilized in the mapping.

The third major type of analysis (in addition to the semivariogram and kriging analyses) that is included in BATGAM© is sequential Gaussian conditional simulation, which can be performed using the **SGSIM** module. This technique can provide a powerful tool, beyond just kriging and the kriging variance, for assessing the uncertainty in estimated maps. In terms of output, SGSIM generates a grid of predicted values similar in format to the kriging output. However, SGSIM generates a large number of equi-probable grids to assess uncertainty, rather than relying on the single 'expected value' grid of kriging. As such, the multiple grids of SGSIM output are viewed via a different visualization module from KT3D Mapping (available with the software at the poster presentation). SGSIM input parameters are obtained from the user via four tabs and associated screens. The first tab (Main) specifies the user's file naming selections, characteristics about the input data set, and the simulation grid size and spacing. The second tab (Search) determines the user's strategy for searching the data during simulation. The third tab (Variogram) requests the user's semivariogram model in the same format as KT3D, including the convention to retrieve a file with semivariogram model parameters from an existing file that was written by GAMV Modeling. The fourth tab (Transformation) requests information from the user about data transformation and/or conditioning with an external data file, if either of those options is selected.

As part of the sequential Gaussian simulation approach, it is assumed that the user's input data follow the Gaussian distribution. As such, the user may require data modules for transforming his/her data into, and out of, a Gaussian distribution (SGSIM also has options for doing automatic normal score transforms). These functions are provided in BATGAM© via the **NSCORE** and **BACKTR** modules. NSCORE performs the Gaussian transformation via what might be viewed as a graphical procedure where the original input data are equated with their normal-score equivalents (i.e., standard normal distribution quantiles). BACKTR then uses this same relationship between standard Gaussian values and input data values to back-transform the grid of simulated values into the user's original data units.

One additional module included in BATGAM© is **GAM** which is used, similar to GAMV, to calculate experimental semivariogram values from data. However, in the case of GAM, the data are simulated values generated by SGSIM. As such, there is generally a large number of these simulated data, and they are located upon a well defined grid. The program GAM is optimized to calculate semivariograms under these conditions, whereas the program GAMV might require extensive execution time. Output from both GAM and GAMV can be viewed with the GAMV Modeling module.

3 Conclusion

BATGAM© is a modest but useful software product which implements some of GSLIB's semivariogram, kriging, and conditional simulation modules. Battelle has recently decided to make it available free of charge, in the hope that others may find it useful. We plan to enhance and expand BATGAM© in the future, depending on Battelle's needs and positive feedback from the geostatistical community.

Reference

Deutsch, C.V., and A.G. Journel, (1998), GSLIB Geostatistical Software Library and User's Guide, Oxford University Press, New York and Oxford, 369 pp.

INDEX

Quantitative Geology and Geostatistics

1. F.M. Gradstein, F.P. Agterberg, J.C. Brower and W.S. Schwarzacher: *Quantitative Stratigraphy*. 1985 ISBN 90-277-2116-5
2. G. Matheron and M. Armstrong (eds.): *Geostatistical Case Studies*. 1987 ISBN 1-55608-019-0
3. *Cancelled*
4. M. Armstrong (ed.): *Geostatistics.* Proceedings of the 3rd International Geostatistics Congress, held in Avignon, France (1988), 2 volumes. 1989 Set ISBN 0-7923-0204-4
5. A. Soares (ed.): *Geostatistics Tróia* '92, 2 volumes. 1993 Set ISBN 0-7923-2157-X
6. R. Dimitrakopoulos (ed.): *Geostatistics for the Next Century*. 1994 ISBN 0-7923-2650-4
7. M. Armstrong and P.A. Dowd (eds.): *Geostatistical Simulations.* 1994 ISBN 0-7923-2732-2
8. E.Y. Baafi and N.A. Schofield (eds.): *Geostatistics Wollongong* '96, 2 volumes. 1997 Set ISBN 0-7923-4496-0
9. A. Soares, J. Gómez-Hernandez and R. Froidevaux (eds.): *geoENV I - Geostatistics for Environmental Applications*. 1997 ISBN 0-7923-4590-8
10. J. Gómez-Hernandez, A. Soares and R. Froidevaux (eds.): *geoENV II - Geostatistics for Environmental Applications*. 1999 ISBN 0-7923-5783-3
11. P. Monestiez, D. Allard and R. Froidevaux (eds.): *geoENV III - Geostatistics for Environmental Applications*. 2001 ISBN 0-7923-7106-2; Pb 0-7923-7107-0
12. M. Armstrong, C. Bettini, N. Champigny, A. Galli and A. Remacre (eds.): *Geostatistics Rio 2000*. 2002 ISBN 1-4020-0470-2
13. X. Sanchez-Vila, J. Carrera and J.J. Gómez-Hernández (eds.): *geoENV IV - Geostatistics for Environmental Applications.* 2004 ISBN 1-4020-2007-4; Pb 1-4020-2114-3
14. O. Leuangthong and C.V. Deutsch (eds.): *Geostatistics Banff 2004*, 2 volumes. 2005 Set ISBN 1-4020-3515-2

springeronline.com