ELSEVIER
JAI

APPLICATIONS OF MANAGEMENT SCIENCE
VOLUME 12

APPLICATIONS OF MANAGEMENT SCIENCE: IN PRODUCTIVITY, FINANCE, AND OPERATIONS

KENNETH D. LAWRENCE
RONALD K. KLIMBERG
Editors

APPLICATIONS OF MANAGEMENT SCIENCE: IN PRODUCTIVITY, FINANCE, AND OPERATIONS

APPLICATIONS OF MANAGEMENT SCIENCE: IN PRODUCTIVITY, FINANCE, AND OPERATIONS.

Series Editor: Kenneth D. Lawrence

Recent Volumes:

Volume 8:	Network Optimization in Applications
Volume 9:	Engineering Applications
Volume 10:	Multi-Criteria Applications
Volume 11:	Mathematical Programming

APPLICATIONS OF MANAGEMENT SCIENCE VOLUME 12

APPLICATIONS OF MANAGEMENT SCIENCE: IN PRODUCTIVITY, FINANCE, AND OPERATIONS

EDITED BY

KENNETH D. LAWRENCE
New Jersey Institute of Technology, USA

AND

RONALD K. KLIMBERG
Saint Joseph's University, USA

ELSEVIER
JAI

Amsterdam – Boston – Heidelberg – London – New York – Oxford
Paris – San Diego – San Francisco – Singapore – Sydney – Tokyo
JAI Press is an imprint of Elsevier

JAI Press is an imprint of Elsevier
The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB, UK
Radarweg 29, PO Box 211, 1000 AE Amsterdam, The Netherlands
525 B Street, Suite 1900, San Diego, CA 92101-4495, USA

First edition 2006

Notice
No responsibility is assumed by the publisher for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions or ideas contained in the material herein. Because of rapid advances in the medical sciences, in particular, independent verification of diagnoses and drug dosages should be made

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

ISBN-13: 978-0-7623-1221-4
ISBN-10: 0-7623-1221-1
ISSN: 0276-8976 (Series)

Printed and bound in The Netherlands

06 07 08 09 10 10 9 8 7 6 5 4 3 2 1

CONTENTS

PART I: DATA ENVELOPMENT ANALYSIS

PART II: MULTI-CRITERIA APPLICATIONS

PART III: OPERATIONAL APPLICATIONS

PART IV: FINANCIAL AND OTHER APPLICATIONS

LIST OF CONTRIBUTORS

Allan Ashley	School of Business, Adelphi University, Garden City, NY, USA
David L. Bakuli	Department of Economics and Management, Westfield State College, MA, USA
Yong Soo Chun	School of Business Administration, Inha University, Inchon, Korea
William W. Cooper	University of Texas, Austin, TX, USA
Steven Cosares	Frank G. Zarb School of Business, Hempstead, NY, USA
Peter M. Ellis	Department of Business Administration, Utah State University, Logan, UT, USA
Jerry Fjermestad	School of Management, NJIT, Newark, NJ, USA
Surendra M. Gupta	Department of Mechanical and Industrial Engineering, Northeastern University, Boston, MA, USA
Karen M. Hogan	Department of Finance, Haub School of Business, Saint Joseph's University, Philadelphia, PA, USA
Zhimin Huang	School of Business, Adelphi University, Garden City, NY, USA
Prasit Imtanavanich	Department of Mechanical and Industrial Engineering, Northeastern University, Boston, MA, USA
Seongho Kim	School of Business Administration, Inha University, Inchon, Korea

Ronald Klimberg	Haub School of Business, Saint Joseph's University, Philadelphia, PA, USA
N. K. Kwak	John Cook School of Business, Saint Louis University, St. Louis, MO, USA
Kenneth D. Lawrence	School of Management, NJIT, Newark, NJ, USA
Sheila M. Lawrence	College of Business, Rutgers University, NJ, USA
Vedran Lelas	Plymouth State University, Plymouth, NH, USA
Susan X. Li	School of Business, Adelphi University, Garden City, NY, USA
Thomas W. Lin	Elaine and Kenneth Leventhal School of Accounting, University of Southern California, Los Angeles, CA, USA
Hai Lu	Rotman School of Management, University of Toronto, Toronto, ON, Canada
Seamus M. McGovern	Department of Mechanical and Industrial Engineering, Northeastern University, Boston, MA, USA
Hani I. Mesak	Department of Marketing and Analysis, College of Administration and Business, Louisiana Tech University, Ruston, LA, USA
Virginia M. Miori	Haub School of Business, Saint Joseph's University, Philadelphia, PA, USA
Amitava Mitra	Office of the Dean and Department of Management, College of Business, Auburn University, Auburn, AL, USA
Daniel O'Leary	Marshall School of Business, University of Southern California, Los Angeles, CA, USA
Gerard T. Olson	Department of Finance, Villanova University, Villanova, PA, USA

Jayprakash G. Patankar	Department of Management, The University of Akron, Akron, OH, USA
Gary R. Reeves	Department of Management Science, Moore School of Business, University of South Carolina, Columbia, SC, USA
Fred J. Rispoli	Department of Mathematics and Computer Science, Dowling College, Oakdale, NY, USA.
Joseph Sarkis	Graduate School of Management, Clark University, Worcester, MA, USA
David E. Schultz	Department of Engineering, University of Southern Indiana, Evansville, IN, USA
Inshik Seol	Graduate School of Management, Clark University, Worcester, MA, USA
George P. Sillup	Department of Management, Haub School of Business, Saint Joseph's University, Philadelphia, PA, USA
David W. Sullivan	Texas Commission on Environmental Quality, Austin, TX, USA
Suresh K. Tadisina	Department of Management, Southern Illinois University, Carbondale, IL, USA
Hongkai Zhang	School of Business, East Central University, Ada, OK, USA

This page intentionally left blank

DEDICATION

This peer-reviewed volume is part of an annual series, dedicated to the presentation and discussion of state-of-the-art studies in the application of management science to the solution of significant managerial decision-making problems. It is hoped that this research annual will significantly aid in the dissemination of actual applications of management science in both the public and private sectors. Volume 1 is directed toward the applications of mathematical programming to (1) multi-criteria decision making, (2) supply chain management, (3) performance management, and (4) risk analysis. Its use can be found both in the university classes in management science and operations research (management and engineering schools), as well as to both the researcher and practitioner of management science and operations research. Series information at: http://www.elsevier.com/locate/series/mansc

This page intentionally left blank

EDITORIAL BOARD

This page intentionally left blank

PART I:
DATA ENVELOPMENT ANALYSIS

This page intentionally left blank

MEASURING ORGANIZATIONAL EFFICIENCY WITH DATA ENVELOPMENT ANALYSIS: THE HORACE MANN INSURANCE COMPANY

Peter M. Ellis

ABSTRACT

Data envelopment analysis is used here to track organizational efficiency over time. Input and output values are discounted with the consumer price index for equity. The model obtains annual comparisons of efficiency by obtaining outputs from a designated set of inputs.

Entities, whether governmental, private or commercial, can be thought of as having a set of inputs, some processing activities and a set of outputs. There is a sense that the entity is efficient if it obtains a great amount of output while expending few inputs. An efficiency analysis is a study to measure the relation between inputs and outputs. The study will be carried out here with data envelopment analysis (DEA). DEA is typically used to compare the relative efficiency of each of a set of operating units. These are usually called

Applications of Management Science: In Productivity, Finance, and Operations
Applications of Management Science, Volume 12, 3–11

ISSN: 0276-8976/doi:10.1016/S0276-8976(06)12001-5

decision-making units (DMUs). The technique was pioneered by Charnes and Cooper in the 1970s. Two excellent works are those of Charnes, Cooper, and Rhodes (1978) and Banker, Charnes, and Cooper (1984). Cooper, Seiford, and Tone (1999) also recently published a book on the use of DEA.

The technique has been applied in many environments. Works from the public sector include McCarty and Yaisawarng (1993) and Vanden Eeckaut, Tulkens, and Jamar (1993). Applications in the financial sector include the works by Lovell and Pastor (1997), Barr, Seiford, and Siems (1993) and Siems and Barr (1998).

DEA has traditionally been used to establish a relative measure of efficiency on any of the several competing entities. It will be used here instead to track efficiency of a single entity over time. The operating statistics at consecutive time points are used as the distinct entities. This application of DEA is believed to be unique. The technique will be illustrated here for the property–casualty insurance industry. Several input and output variables will be selected.

DEA has been used in a property–casualty insurance industry study by Cummins, Weiss, and Zi (1999). Cummins and Zi (1998) also used the method to compare life insurers. These works have all been applied cross-sectionally, where several DMUs are compared at a given time point.

The work done here proposes to track the efficiency of a single organization over time instead of comparing efficiency of several competing organizations at a single time point. The organization being considered is the Horace Mann Group of property–casualty insurers. The Horace Mann Group was founded in 1945 and has stayed close to its original purpose of providing individual automobile and homeowners' insurance coverage to people with ties to education.

DATA ENVELOPMENT ANALYSIS

The parameters and variables of DEA are:

n = number of DMUs under comparison
E_i = efficiency of DMU$_i$
u_{ij} = input level of DMU$_i$ for input measure j
v_{ij} = output level of DMU$_i$ for output measure j
w_i = percentage weight of DMU$_i$ in the weighted average composite DMU.

DEA seeks to construct a weighted average of the several DMUs being compared. The required weighted average is one that has the greatest efficiency. This embraces two things. First, the weighted average will have outputs that are at least as great as those of any individual DMU. Second, the weighted average will use input levels that are at most only E_i percent as large as those of any DMU$_i$. It may turn out that the entire weighted average composite is formed from just one most efficient DMU, or perhaps the efficient composite is formed from a larger subset or the entirety of all DMUs.The formal DEA model for DMU$_i$ is:

$$\text{Minimize} \quad z = E_i \tag{Ia}$$

$$\text{Subject to} \quad \sum_{i=1}^{n} w_i = 1 \tag{Ib}$$

$$\sum_{i=1}^{n} v_{ij} w_i \geq v_{ij} \quad \text{all outputs } j \tag{Ic}$$

$$\sum_{i=1}^{n} u_{ij} w_i \leq u_{ij} E_i \text{ all inputs } j \quad \text{all } w_i, \;\; E_i \geq 0 \tag{Id}$$

The objective function in (Ia) serves to determine the relative efficiency level E_i for DMU$_i$. Constraint (Ib) forces the relative weights in the composite DMU to sum to one. This causes the individual weights to be percentages of the total composite entity. The constraints of (Ic) force the several outputs of the composite to be at least as large as the corresponding outputs from any individual DMU, which clearly is needed for an efficient entity. Finally, constraints (Id) require that the several inputs of the weighted average composite DMU be no greater than E_i percent of those of any DMU$_i$. As always in a linear programming formulation, the variables are also required to be nonnegative.

The Horace Mann Insurance Group

The Horace Mann Insurance companies is a group consisting of Allegiance Insurance Company, Horace Mann Insurance Company and Teachers Insurance Company. Horace Mann Insurance was founded in 1945 in Springfield, Illinois. The group has for many years specialized in marketing personal lines of insurance to customers with ties to the nation's education system. The group believes that it has a competitive advantage over other similar insurers because of the quality of the pool of insureds. The limitation to educators is believed to yield a risk pool with superior education, judgement and personal values. The lines are principally personal automobile and homeowners insurance. The group has reported operating and investment profits quite consistently for years. It stands apart from the aggregate of the property–casualty insurance industry in that its underwriting activity has generally been profitable each year, while the industry generally has operated with annual losses. The industry aggregate has used investment income to offset operating losses, while Horace Mann has nearly always shown annual gains from both operations and investment activity. Because annual underwriting activities have not experienced the great fluctuations in loss experience and operating profit or loss of the industry aggregate, it is not expected that the efficiency magnitudes will vary widely from year to year. The Horace Mann Group has experienced consistent and steady operating results compared to the industry aggregate. The annual efficiency value will be useful in demonstrating the group's reaction to environmental changes that create variations in factors that affect settlement costs. These include cost inflation in housing, medical care and auto repair.

The first input variable used here is the real value of total assets at the beginning of the year. This is used because total assets are predominantly dedicated to the investment portfolio, consisting of stocks and bonds. This is a statutory requirement in all states, and is designed to keep insurers sufficiently liquid to meet loss payment obligations. Thus, total assets is the surrogate for the investment portfolio. The second input variable is the real value of policyholder surplus. 'Policyholder surplus' is the industrial term for capital plus surplus. It is used because there is a statutory link between the size of policyholder surplus and annual premium volume. Firms are not permitted to write a large volume of insurance coverage from a small capital base. Policyholder surplus thus dictates the volume of insurance business that can be accepted. The final input variable is real operating expenses. This is the cost of all the activities that are borne while operating the insurance enterprise, primarily consisting of marketing or placement, underwriting and loss adjustment.

Real values of these variables are used in order to make them comparable over the entire time horizon of the study. The values are obtained by dividing the nominal value by the urban household consumer price index (CPI) during the years 1981–1982 = 100.

The first output variable is the real value of the loss and loss adjustment expense reserve. It might be initially thought that the expected output would be annual premium volume. However, that has generally been abandoned by researchers because premium volume is set by multiplying unit premium by the number of insureds, so an annual increase in total volume cannot be easily attributed to either a price increase or a wider placement. The insurance enterprise exists to cover loss experience within the risk pool, so the product is really a promise to cover liabilities that arise. The theoretically correct way to calculate this is to add up the discounted value of all future losses that are covered by a policy that is in effect during the year. This is generally not readily available from individual insurers. However, when a loss is reported, the insurer incurs a liability that is shown in the loss reserve liability account, so this serves well as an output measure.

Real underwriting gain or loss is an obvious output variable. This is the annual profit that arises from insurance operations. Similarly, the third output variable is real net investment income. As premium dollars are received, they are added to the investment portfolio. The portfolio has typically been more consistently and steadily profitable than has underwriting activity. For the aggregate property–casualty industry net investment income has been consistently profitable, while underwriting has been consistently unprofitable. The Horace Mann companies have more commonly had profits from both underwriting and investments.

The DMUs in this study are the annual values for the six variables described above. The years used were 1994–2000, being seven in all. Another DMU was introduced to circumvent a common problem in DEA. It is that the output of the linear programming model for any particular DMU will show a single DMU or a subset of all the DMUs as the efficient set. Then the comparable linear programming model is to be run for other DMUs as well. It is very likely that the efficient sets of DMUs from these successive models will not be identical. That is, individual DMUs are being compared with relative efficiency based upon differing standards of comparison. This problem is avoided by introducing the new DMU IDEAL. This was created for the purpose of providing a DMU which is certain to be the efficient set all by itself for comparison to all of the individual DMUs. With IDEAL present, all existing DMUs are being compared to IDEAL and not to each other. This provides a single standard of comparison for all DMUs. IDEAL is formed by assigning to it the minimum values of all the input variables and the maximum values of all the output variables. IDEAL thus will have the greatest output and use the smallest inputs. It will automatically be the efficient DMU that will be used as the standard of comparison for each

Table 1. Inputs and Outputs for the Horace Mann Insurance Group (Efficiency Study).

Year	Inputs ($000)		
	Real Beginning Total Assets	Real Beginning Policyholder Surplus	Real Operating Expenses
2000	356605.6	115639.3	59593.20
1999	373536.6	114519.2	58838.44
1998	392136.8	114888.3	57749.03
1997	452568.8	140525.8	55370.25
1996	460518.1	143062.4	52814.86
1995	482348.4	134962.5	52721.52
1994	507865.7	134093.1	53276.96
IDEAL	356605.6	114519.2	52721.52

Year	Outputs ($000)		
	Real Loss and Loss Adjustment Reserves	Real Underwriting Gain or Loss	Real Net Investment Income
2000	145087.1	−18688.1	20735.19
1999	141292.3	10074.42	22226.29
1998	149103.6	16353.37	23882.82
1997	167793.1	22773.2	26002.49
1996	195251.7	15066.28	29578.71
1995	242495.4	16633.85	32068.24
1994	256551.9	14825.91	30598.51
IDEAL	256551.9	22773.2	32068.24

individual DMU. The real input and output values for the several years of the study and also IDEAL are shown in Table 1. The data was obtained from Best's Aggregates and Averages (various years).

Listing 1 shows the actual linear programming formulation of the DEA problem for the year 1994. The objective function seeks to minimize the efficiency of the group in 1994. If that minimum value turns out to be $E1994 = 1$, then the year 1994 will be found to be fully efficient in comparison to all the years included. The first constraint of Listing 1 is equation (2). It requires that the total of the relative weights of the DMUs add to 1. These weights assign to each year (DMU) its percentage contribution to the efficient set of all DMUs. Because of the inclusion of IDEAL, the weights assigned to the individual years will all be zero, and the weight of IDEAL will equal to 1.

```
MIN    E1994
SUBJECT TO
    2)  W1994 + W1995 + W1996 + W1997 + W1998 + W1999 + W2000 + WIDEAL
   =   1
    3)  256551.9 W1994 + 242495.4 W1995 + 195251.7 W1996
   + 167793.1 W1997 + 149103.6 W1998 + 141292.3 W1999 + 145087.1 W2000
   + 256551.9 WIDEAL >=   256551.9
    4)  14825.91 W1994 + 16633.85 W1995 + 15066.28 W1996 + 22773.2 W1997
   + 16353.37 W1998 + 10074.42 W1999 - 18688.1 W2000 + 22773.2 WIDEAL
   >=   14825.91
    5)  30598.51 W1994 + 32068.24 W1995 + 29578.71 W1996
   + 26002.49 W1997 + 23882.82 W1998 + 22226.29 W1999 + 20735.19 W2000
   + 32068.24 WIDEAL >=   30598.51
    6) - 507865.7 E1994 + 507865.7 W1994 + 482348.4 W1995
   + 460518.1 W1996 + 452568.9 W1997 + 392136.9 W1998 + 373536.6 W1999
   + 356605.6 W2000 + 356605.6 WIDEAL <=   0
    7) - 134093.1 E1994 + 134093.1 W1994 + 134962.5 W1995
   + 143062.4 W1996 + 140525.8 W1997 + 114888.3 W1998 + 114519.2 W1999
   + 115639.3 W2000 + 114519.2 WIDEAL <=   0
    8) - 53276.96 E1994 + 53276.96 W1994 + 52721.52 W1995
   + 52814.86 W1996 + 55370.25 W1997 + 57749.03 W1998 + 58838.44 W1999
   + 59593.2 W2000 + 52721.52 WIDEAL <=   0
END
```

Listing 1. The Linear Programming Formulation: The Horace Mann Insurance Group (1994 Data).

The next three constraints of Listing 1 are shown in equations (3)–(5). These constraints require that the outputs of the efficient DMU IDEAL be at least as great as those of the three outputs for 1994. The last three constraints in Listing 1 require that the inputs of the ideal be no greater than E1994 times the corresponding input values of 1994. It is seen that the efficient set uses the smallest inputs and yields the greatest outputs.

Table 2 shows the optimal solution to the DEA analysis. The overall finding is that years 1995, 1999 and 2000 were perfectly efficient. In 1998, the efficiency was 0.9967 and in 1996 it was 0.9982. Therefore, in every year except 1997 the efficiency was at or nearly 1.0. However, in 1997 it slipped to just 0.9521. Even so, the firm can appreciate that it has remained consistently efficient over this span of time. Efficiency has not swung wildly up or down, nor has it trended up or down. This demonstrates consistent usage of available resources in generating outputs, which indicates the presence of stable managerial practice and control.

The slack entries in Table 2 help to illustrate why any year is not efficient. The slack values are the amounts by which the inputs exceed those needed to be just E2004 times the input level needed to maintain the efficiency level.

Table 2. Optimal DEA Results (Slack Amounts $000).

Year	Efficiency	Real Total Asset Slack	Real Total Policyholder Surplus Slack	Real Operating Expense Slack
2000	1.0000	0	1120.094000	680071.680000
1999	1.0000	0	0	0
1998	0.9967	34271.430000	0	4841.985000
1997	0.9521	74313.910000	19284.30000	0
1996	0.9982	0	11528.57000	0
1995	1.0000	0	0	0
1994	0.9895	145965.30000	18175.90000	0

For 1997, the level of real assets exceeded the needed level by $74313(000). The excess level of policyholder surplus (capital plus surplus) was $19284.3(000). There was zero excess operating expenditure in 1997. The year 2000 is interesting in that it had an efficiency of 1, but still was seen to have slacks in both real policyholder surplus and real operating expense. However, it had the smallest real total asset base upon which satisfactory output levels were realized.

No year had positive slack values for all three inputs and no input had continuing slack for more than two consecutive years. This suggests that Horace Mann management has been diligent in overseeing asset and expense management. Management will conclude that the firm was not run as efficiently in 1997 when compared to other years, but that within two years it had recovered to full efficiency. Further, all of the efficiencies were at least 95%. This shows that the group has produced overall operating results that are fairly consistent over the years.

CONCLUSION

DEA has been applied to the evaluation of comparative annual performance of an organization. The analysis requires selecting sets of input and output measures for each year of the time horizon. It is generally argued that efficiency consists of obtaining relatively greater output while expending relatively fewer inputs. DEA formalizes this with a linear programming formulation that identifies an efficient subset of DMUs. That approach is modified here so that the DMUs are the successive years of the time horizon. Input and output variables are stated in real values. These are obtained by dividing the nominal value by the CPI.

An application of the DEA model was developed here using data from the Horace Mann Insurance Group. The time horizon was the years 1994–2000. It was shown that the Horace Mann Group was relatively efficient in the years 1995, 1999 and 2000. The maximum possible efficiency level for any year is 1.0, and actual results found the individual efficiencies all to be at least 0.95.

Management can use the results of this efficiency analysis to review its ongoing success in asset and expense management. Inputs with slack capacity are those, which have not been fully utilized in generating the designated outputs. Management will therefore find its interest gravitating to these slack values, so that the set of inputs might be satisfactorily deployed.

REFERENCES

Banker, R. D., Charnes, A., & Cooper, W. W. (1984). Some models for estimating technical and scale inefficiencies in data envelopment analysis. *Management Science*, *30*, 1078–1092.

Barr, R. S., Seiford, L. M., & Siems, T. F. (1993). An envelopment-analysis approach to measuring the managerial efficiency of banks. *Annals of Operations Research*, *45*, 1–19.

Best, A.M. Company. (various years). *Aggregates and averages*. Oldwick, NJ.

Charnes, A., Cooper, W. W., & Rhodes, E. (1978). Measuring the efficiency of decision making units. *European Journal of Operational Research*, *2*, 429–444.

Cooper, W. W., Seiford, L. M., & Tone, K. (1999). *Data envelopment analysis*. Boston: Kluwer Academic.

Cummins, J. D., & Zi, H. (1998). Comparison of frontier efficiency methods: An application to the U.S. Life Insurance Industry. *Journal of Productivity Analysis*, *10*(2), 131–152.

Cummins, J. D., Weiss, M. A., & Zi, H. (1999). Organizational form and efficiency: The coexistence of stock and mutual property-liability insurers. *Management Science*, *45*, 1254–1269.

Lovell, C. A. K., & Pastor, J. T. (1997). Target setting: An application approach to a bank branch network. *European Journal of Operational Research*, *98*, 290–299.

McCarty, T. A., & Yaisawarng, S. (1993). Technical efficiency in New Jersey school districts. In: H. O. Fried, C. A. K. Lovell & S. S. Schmidt (Eds), *The measurement of productive efficiency*. New York: Oxford University Press.

Siems, T.F., & Barr, R.S. (1998). *Benchmarking the productive efficiency of U.S. banks*. Financial Industry Studies Dallas (pp. 11–24). Federal Reserve Bank of Dallas.

Vanden Eeckaut, P., Tulkens, H., & Jamar, M. A. (1993). Cost efficiency in Belgian municipalities. In: H. O. Fried, C. A. K. Lovell & S. S. Schmidt (Eds), *The measurement of productive efficiency*. New York: Oxford University Press.

This page intentionally left blank

CONSUMER-PACKAGED FREIGHT DELIVERY EVALUATION: A MULTIPLE CRITERIA DATA ENVELOPMENT ANALYSIS APPROACH

Gary R. Reeves, Kenneth D. Lawrence and Sheila M. Lawrence

ABSTRACT

This research deals with the evaluation of the efficiency of consumer freight package delivery. Model inputs include number of delivery and administrative employees, labor hours, operating costs, and number of deliver vehicles. Outputs include number of packages delivered, percent on-time, percent lost, percent damaged, revenue per package, and customer satisfaction. The methodology used to evaluate the efficiency of the decision-making units (DMUs) under consideration is data envelopment analysis (DEA) involving multiple criteria. The inclusion of additional criteria beyond basic DEA efficiency can improve discriminating power between DMUs and also tends to yield more reasonable weights on model inputs and outputs.

Applications of Management Science: In Productivity, Finance, and Operations
Applications of Management Science, Volume 12, 13–21

ISSN: 0276-8976/doi:10.1016/S0276-8976(06)12002-7

1. INTRODUCTION

This paper deals with a multiple criteria analysis of transportation productivity of small package carriers. These include such organizations as the U.S. Postal Service (UPS) and Federal Express. For many shippers, such as electronics firms, catalog merchants, cosmetic companies, and book distributors, these carriers are an extremely important part of the U.S. domestic transportation system.

Productivity improvements are vital to the success of any logistics system. Shippers and carriers can achieve productivity gains through increased efficiency and effectiveness. Productivity improvements may be achieved in areas such as:

- systems and design (local pickup and delivery, inbound consolidation);
- use of equipment and labor (monitoring and tracking systems, fuel efficiency, flexible pick-up and delivery hours, compensation incentives);
- performance of labor and equipment (i.e., compensation based on productivity, fuel efficiency).

From a shipper's viewpoint, measures of effectiveness and efficiency include:

- Claims ratios
- Damage ratios
- Variability in transit times
- On-time pick-up
- On-time delivery
- Cost per ton mile
- Billing accuracy
- Customer complaint frequency

From a carrier's viewpoint, measures include:

- monetary contribution by traffic lane;
- monetary contribution by shipper;
- monetary contributor by sales person;
- monetary contribution by terminal.

Such measures are used for internal performance evaluation. The measurement process begins with data collection in terms of monetary values. This data is segmented by the firm's functional units into a time series. Cost data is compared to some type of macro output as a percent of sales (i.e., total transportation cost as a percent of sales). Typically outbound transportation is measured and then inbound freight cost and flows are measured. Next, physical measures and budgets are formed for each transportation activity (i.e., weight order, miles, etc.) and are tracked with the various activities over various time intervals. Then physical units can be measured against labor and non-labor costs to track cost per mile or per ton mile.

Next, a set of operational goals is set for the overall operation. These goals can be in terms of unit cost. With these, a measure of performance and actual versus standard comparisons can be developed. In a more sophisticated manner, standards for labor and non-labor activities can be developed for each transportation activity. Thus, transportation requirements can be converted to vehicle load or dollars of cost. Thereupon, performance measurement of labor and non-labor inputs by activities against variables are made.

Finally, a system of productivity measurement can be developed. In this analysis, the performance data is merged with financial data to provide an overall view of the operation. Given this measurement system, the firm is in a position to test alternatives and seek trade-offs with present operations. Productivity measurement allows the firm's decision makers to plan and to control their transportation system.

2. DATA ENVELOPMENT ANALYSIS

Data Envelopment Analysis (DEA) was developed by Charnes, Cooper, and Rhodes (1978), as a linear programming based methodology for evaluating the efficiency of decision making units (DMUs) either within a company or across an industry. A DEA model for evaluating the efficiency of a particular DMU (DMU$_0$) can be stated as:

$$\max h_0 = \sum_{r=1}^{s} u_r y_{rj_0}$$

subject to

$$\sum_{r=1}^{s} v_i x_{ij_0} = 1$$

$$\sum_{r=1}^{s} u_r y_{rj} - \sum_{i=1}^{m} v_i x_{ij} \leq 0, \quad j = 1, \ldots n$$

$$u_r, v_i \geq 0, \quad \text{for all } r \text{ and } i \tag{1}$$

where there are n different DMUs under consideration ($j = 1, \ldots, n$), s output measures ($r = 1, \ldots, s$) and m input measures ($i = 1, \ldots, m$) for each DMU. Also, y_{rj} is the value of the rth output for the jth DMU, x_{ij} the value of the ith input for the jth DMU, u_r the weight given to the rth output, v_i the weight given to the ith input, and h_0 the relative efficiency of DMU_0, the DMU under evaluation.

DMU_0 is efficient if and only if the maximum value of h_0 is equal to 1. Model (1) is solved for each DMU. Decision makers can use these efficiency ratings to identify those DMUs, which are in need of improvement. A survey of DEA models and applications is contained in Charnes, Cooper, Lewin, and Seiford (1994).

3. MULTIPLE CRITERIA DEA

DEA has been applied in a wide variety of problems in both the public and private sectors, however, some concerns have been raised in certain situations about the discriminating power of DEA and the reasonableness of the values of some DEA model input and output weights. With regard to discriminating power, DEA models sometimes classify too large a proportion of DMUs as efficient without providing any further mechanism for differentiating between these DMUs. In addition, the input and/or output weights generated by the model that resulted in a DMU being classified as efficient may be unreasonable or undesirable in practice. Motivated at least in part by these issues, several researchers have begun to explore the relationship or interface between DEA and multiple criteria decision-making (MCDM). They include Sexton, Silkman, and Hogan (1986), Golany (1988), Belton and Vickers (1993), Doyle and Green (1993), Stewart (1996), Joro, Korhonen, and Wallenius (1998), Li and Reeves (1999), and Sarkis (2000).

To facilitate the transition to a multiple criteria model, the single criteria DEA model (1) can be expressed equivalently through the introduction of deviational variables:

$$\min\ d_0 \quad (\text{or } \max\ h_0 = \sum_{r=1}^{s} u_r y_{rj_0})$$

subject to

$$\sum_{r=1}^{s} v_i x_{ij_0} = 1$$

$$\sum_{r=1}^{s} u_r y_{rj} - \sum_{i=1}^{m} v_i x_{ij} + d_j = 0, \quad j = 1, \ldots, n$$

$$u_r, v_i, d_j \geq 0, \quad \text{for all } r,\ i \text{ and } j \tag{2}$$

where d_0 is the deviational variable for DMU$_0$ and d_j the deviational variable for the *j*th DMU. DMU$_0$ efficient if and only if the minimum value of d_0 equal to 0 and inefficient otherwise.

While the value of d_0 is a good measure of DMU$_0$'s inefficiency, it is not the only possible measure. Other measures have been suggested in previous studies, most of which are functions of deviational variables. Sexton et al. (1986) used the weighted sum of all deviational variables as the objective function in their DEA model, while Stewart (1996) discussed minimizing the maximum deviational variable. Li and Reeves (1999) used all three of these criteria, minimizing d_0, minimizing the maximum deviation (minmax), and minimizing the sum of the deviations (minsum), in their multiple criteria DEA model:

$$
\begin{aligned}
&\min\ d_0 \quad (\text{or } \max\ h_0 = \sum_{r=1}^{s} u_r y_{rj_0})\\
&\min\ M\\
&\min \sum_{j=1}^{n} d_j\\
&\text{subject to}\\
&\sum_{i=1}^{m} v_i x_{ij_0} = 1\\
&\sum_{i=1}^{m} u_r y_{rj} - \sum_{i=1}^{m} v_i x_{ij} + d_j = 0, \quad j = 1, \ldots, n\\
&M - d_j \geq 0, \quad j = 1, \ldots, n\\
&u_r, v_i, d_j \geq 0, \quad \text{for } r, i, \text{and } j
\end{aligned}
\tag{3}
$$

The first objective of this model is identical to the single objective of maximizing the efficiency, or equivalently minimizing the inefficiency, of DMU$_0$ in models (1) and (2), respectively. The M variable in the second objective takes on the maximum value of all the deviational variables, while the third objective represents the sum of the deviational variables, where both are considered measures of inefficiency to be minimized. The additional constraints, $M{-}dj{\geq}0$ ($j = 1, \ldots, n$), force M to equal or exceed the maximum individual deviation; but do not alter the set of allowable input and output weights.

These additional criteria (minmax and minsum) in model (3) do not give favorable consideration to the DMU under evaluation, as the original DEA criterion does. Therefore, efficiencies defined under these criteria are more restrictive than in classical DEA, making it more difficult for a DMU to achieve minmax or minsum efficiency. As a result, the minmax and minsum criteria generally yield fewer efficient DMUs. Including these additional criteria, therefore, should result in improved discriminating power. Also, since these criteria are functions of deviational variables and each deviational variable is related to a particular input–output constraint, minimizing M or $\sum d_j$ should serve to limit the flexibility in the weights assigned to model inputs and outputs.

4. AN EXAMPLE

To illustrate, a consumer packaged freight delivery example involving 15 DMUs, 5 model inputs and 6 model outputs will be presented. Model inputs include number or employees (delivery and administrative), number of delivery vehicles, operating costs, and labor hours. Model outputs include percent of deliveries on-time, lost, or damaged, number of packages, revenue per package, and customer satisfaction. The data on inputs and outputs is shown in Tables 1 and 2.

Table 1. Outputs.

DMU	Percent On-Time Satisfaction	♯ of Packages	Percent Lost	Percent Damaged	Revenue/ Package	Customer
1	99	7240	1	2	16.05	3.1
2	98	6840	1	3	12.95	8.6
3	96	9782	2	1	18.40	7.4
4	97	10141	2	1	25.00	9.4
5	95	7603	3	2	18.07	9.6
6	99	9423	1	2	19.06	8.5
7	99	8844	1	1	17.02	8.8
8	99	7187	2	2	18.14	9.2
9	98	9741	2	3	22.12	9.3
10	96	11507	3	2	16.04	9.5
11	97	10094	2	1	19.08	9.6
12	98	8473	2	2	20.40	7.7
13	95	9242	1	1	17.09	8.3
14	95	9675	2	2	18.84	9.2
15	99	8833	3	3	23.45	7.5

Table 3 indicates the value of the inefficiency score (d_0) for each DMU when each of the objectives of model (3) ($d_0, M, \sum_{j=1}^{n} d_j$) are minimized. The values in the d_0 column represent the traditional DEA scores. For this data

Table 2. Inputs.

DMU	Number of Delivery Employees	Operating Costs	Labor Hours	Number of Delivery Vehicles	Number of Administrative Employees
1	30	2500	1100	10	6
2	33	2604	1400	9	5
3	27	2209	984	8	3
4	44	3432	1700	11	7
5	19	1821	809	12	3
6	24	2600	941	9	4
7	19	1547	884	7	3
8	26	2304	1017	12	3
9	31	2591	1381	10	4
10	32	2482	1241	9	3
11	24	2200	987	12	4
12	27	2253	1107	11	3
13	28	2470	1081	9	3
14	26	2581	1094	12	3
15	19	1691	902	13	3

Table 3. Performance Scores.

DMU	d_0	M	$\sum_{j=1}^{n} d_j$
1	0.0702	0.2196	0.3353
2	0	0.2209	0.1268
3	0	0.0029	0.0444
4	0.0425	0.1811	0.3190
5	0	0.0490	0
6	0	0.0929	0.0076
7	0	0	0
8	0	0.1890	0.1432
9	0	0	0
10	0	0	0
11	0	0.1606	0.1391
12	0.0101	0.1552	0.1270
13	0.0326	0.2263	0.2087
14	0	0.1900	0.1710
15	0	0.0594	0

set, 11 of the 15 DMUs (or 73%) were classified as efficient (all but DMUs 1, 5, 12, and 13). The inclusion of the minsum and minmax criteria, however, reduce the number of efficient DMUs to 5 and 3, respectively, providing additional discrimination.

The inclusion of the additional criteria effectively increases the number of categories of DMUs from 2 (efficient, inefficient) to 4. The four categories from most to least desirable being: (1) efficient across all 3 criteria – DEA, minsum, and minmax; (2) DEA and minsum efficient, but not minmax; (3) DEA efficient, but neither minsum or minmax; and (4) inefficient.

Only three DMUs, 7, 9, and 10, are efficient with respect to all three criteria. An additional two DMUs, 5 and 15, are DEA and minsum efficient, but not minmax efficient. Six DMUs are DEA efficient only. The classification of the final four DMUs originally identified as inefficient (1, 4, 12, 13) remains unchanged.

Further, the additional criteria allow for additional discrimination within categories. In category 2, DMUs 5 and 15 are both DEA and minsum efficient, but DMU 5 has a lower minmax score. Of the six DMUs in category 3 that are DEA efficient only (2, 3, 6, 8, 11, 14), DMUs 3 an 6 dominate the remaining four DMUs with respect to both the minsum and minmax criteria. Of the remaining four DMUs in this category, DMU 11 dominates DMUs 8 and 14 and DMU 8 dominates DMU 14.

5. CONCLUSIONS

This paper has demonstrated how the inclusion of additional criteria in DEA models can increase their discriminating power between DMUs beyond the basic efficient/inefficient classification. These results were illustrated in the context of an application involving consumer packaged freight delivery.

REFERENCES

Belton, V., & Vickers, S. P. (1993). Demystifying DEA – A visual interactive approach based on multiple criteria analysis. *Journal of the Operational Research Society*, *39*, 725–734.

Charnes, A., Cooper, W. W., Lewin, A. Y., & Seiford, L. M. (1994). *Data envelopment analysis: Theory, methodology and application*. Boston: Kluwer.

Charnes, A., Cooper, W. W., & Rhodes, E. (1978). Measuring the efficiency of decision making units. *European Journal of Operational Research*, *2*, 429–444.

Doyle, J., & Green, R. (1993). Data envelopment analysis and multiple criteria decision making. *Omega*, *21*, 713–715.

Golany, B. (1988). An interactive MOLP procedure for the extension of DEA to effectiveness analysis. *Journal of the Operational Research Society*, *39*, 725–734.

Joro, T., Korhonen, P., & Wallenius, J. (1998). Structural comparison of data envelopment analysis and multiple objective linear programming. *Management Science*, *44*, 962–970.

Li, X., & Reeves, G. R. (1999). A multiple criteria approach to data envelopment analysis. *European Journal of Operational Research*, *115*, 507–517.

Sarkis, I. (2000). A comparative analysis of DEA as a discrete alternative multiple criteria decision tool. *European Journal of Operational Research*, *123*, 543–557.

Sexton, T. R., Silkman, R. H., & Hogan, A. J. (1986). Data envelopment analysis: Critique and extensions. In: R. H. Silkman (Ed.), *Measuring efficiency: An assessment of data envelopment analysis* (pp. 73–105). San Francisco: Jossey-Bass.

Stewart, T. J. (1996). Relationships between data envelopment analysis and multiple criteria decision making. *Journal of the Operational Research Society*, *47*, 654–665.

This page intentionally left blank

EVALUATING THE OPERATIONAL EFFICIENCY OF RAILWAY SYSTEMS WITH SENSITIVITY ANALYSES TO SAMPLING VARIATIONS

N. K. Kwak, Yong Soo Chun and Seongho Kim

ABSTRACT

Data Envelopment Analysis (DEA) is a nonparametric mathematical programming technique used to measure the relative efficiency of the production organization's operations. This paper presents the theoretical measures of the railway systems, along with the bootstrap DEA analysis. A DEA model is applied to evaluate the relative efficiency of railway operations of 29 UIC (Union Internationale des Chemins de fer) countries, based on the data obtained from the International UIC publications. The bootstrap DEA analysis provides information (bias estimates) on the sensitivity of the DEA efficiency index to the sampling variations. The model results are analyzed and evaluated in terms of their relative operational performance efficiency. The model results facilitate an organization's decision-making by providing valuable information.

Applications of Management Science: In Productivity, Finance, and Operations
Applications of Management Science, Volume 12, 23–40

ISSN: 0276-8976/doi:10.1016/S0276-8976(06)12003-9

1. INTRODUCTION

In the global logistics and supply chain management, operational efficiency is an important factor for an organization to survive in an increasingly competitive environment. Since transportation cost comprises a substantial portion of product delivery cost in the global supply chain, the competitiveness depends not only on the efficiency and effectiveness of an organization but also on the efficiency of a country's social overhead capital.

To enhance the operational efficiency of an organization, it should be evaluated in a proper manner. Traditionally, various single ratios (e.g., production amount of output per worker) have been widely used as efficiency indices because they are relatively easy to understand. However, when the operating units of organizations have multiple inputs and outputs, a single ratio measurement does not reflect the total efficiency of an organization, but only the partial factor efficiency. The aggregation of partial factor efficiency is not the total factor efficiency. The total factor efficiency should consider all inputs used and all outputs produced by an organization simultaneously.

Data Envelopment Analysis (DEA) proposed by Charnes, Cooper, and Rhodes (1978) is an application of nonparametric mathematical programming technique. It is arguably the most widely accepted technique to aggregate multiple inputs and outputs into a single efficiency index. Such a widespread acceptance of DEA for efficiency evaluation is due to its well-known merits over traditional approaches (e.g., scoring method, weighted average method).

A plethora of DEA-related studies has appeared in the literature of management science and operations research. Seiford (1996) traced the evolution of DEA and listed over 700 research publications. Recently, other DEA studies have appeared that were not previously covered by Seiford. Kwak and Kim (2000) listed 68 recent DEA application papers in the areas of education, energy, finance, health care, manufacturing, public services, and others. Seiford and Zhu (1998) discussed an acceptance system decision rule, extending the potential use of DEA for credit applicant acceptance systems which was proposed by Troutt, Rai, and Zhang (1996). Kwak, Choi, and Kim (2001) applied DEA to the efficiency evaluation of 98 research university libraries in the United States and Canada. Elkins and Lawrence (2001) illustrated a multiple criteria DEA approach to measure productivity utilizing multi-criteria mathematical programming. Holvad, Hougard, Kronborg, and Kvist (2004), Karlafatis (2004), and Martin, Gutierrez, and

Roman (2004) used the DEA to measure the operational efficiency of various transportation systems.

DEA is a mathematical programming approach for measuring efficiency, which does not include explicitly stochastic variations in the model, so it may be sensitive to such variations. A variety of research efforts has been devoted to incorporate stochastic variations into the DEA model. These efforts can be classified into two sets of studies. The first set focuses on developing a model by direct incorporation of stochastic variations (Cooper, Huang, Lelas, Li, & Olesen, 1998; Land, Lovell, & Thore, 1993; Li, 1998; Olesen & Petersen, 1995), and the second set seeks to incorporate indirectly stochastic variations into the model using resampling methodology, such as jackknife and bootstrapping (Ferrier & Hirschberg, 1997; Gstach, 1995; Simar & Wilson, 1998). The bootstrapping in DEA provides the empirical distribution of the DEA efficiency estimator.

In this paper, we apply DEA to the evaluation of railway operational efficiency of 29 UIC (International Union of Railways) member countries as a part of measurement for social overhead capital efficiency. In addition, we bootstrap DEA efficiency index to obtain the additional information on the sensitivity to the sampling variations. The paper is organized to present the theoretical backgrounds of evaluation methods, followed by a DEA application. The results of the efficiency evaluations are analyzed and interpreted for better understanding.

2. DEA AND THE BOOTSTRAP

2.1. Data Envelopment Analysis

DEA is an application of a mathematical programming technique that is used to measure the relative efficiency of organizational units, referred to as decision-making units (DMUs). The organizational units utilize similar resources (inputs) to generate similar products (outputs). The efficiency of an organization's operations is commonly measured by the ratio of output to input of resources. When the efficiency measure involves multiple items in input and/or output, a set of weights is required to aggregate these items into a single virtual input and/or output. For the purpose of relative efficiency comparisons, a group of entities, which produce the same outputs and use the same inputs, should be used to evaluate each other with each entity having a certain degree of freedom in managerial decision-making. Charnes et al. (1978) suggested the following mathematical programming model to obtain a best set of weights of input and output items for each entity evaluated.

$$
\begin{aligned}
&\text{Maximize}: && \frac{\mathbf{u}^{\mathrm{T}}\mathbf{y}_k}{\mathbf{v}^{\mathrm{T}}\mathbf{x}_k} \\
&\text{Subject to}: && \\
& && \frac{\mathbf{u}^{\mathrm{T}}\mathbf{y}_j}{\mathbf{v}^{\mathrm{T}}\mathbf{x}_j} \leq 1 \text{for}\, j = 1, 2, \ldots, n \\
& && \mathbf{u} \geq \varepsilon \mathbf{1} \\
& && \mathbf{v} \geq \varepsilon \mathbf{1}
\end{aligned}
\tag{1}
$$

where

$\mathbf{y}_j = (s \times 1)$ vector of observed outputs for the j th DMU
$\mathbf{x}_j = (m \times 1)$ vector of observed inputs for the j th DMU
$\mathbf{u} = (s \times 1)$ vector of ouput weight variables
$\mathbf{v} = (m \times 1)$ vector of input weight variables
$\mathbf{1}$ = vector with each component equal to one
n = number of DMUs
s = number of outputs
m = number of inputs
ε = a non-Archimedean small positive number

In model (1), the decision variables are the weight vectors **u**, **v** to maximize the value of the objective function $\mathbf{u}^{\mathrm{T}}\mathbf{y}_k/\mathbf{v}^{\mathrm{T}}\mathbf{x}_k$. The numerator $\mathbf{u}^{\mathrm{T}}\mathbf{y}_k$ and the denominator $\mathbf{v}^{\mathrm{T}}\mathbf{x}_k$ of the objective function are the virtual output and input of DMU k, respectively. Therefore, the objective function value $\mathbf{u}^{\mathrm{T}}\mathbf{y}_k/\mathbf{v}^{\mathrm{T}}\mathbf{x}_k$ represents a relative efficiency score for DMU k. The constraints mean that virtual output-to-input ratio should not exceed 1 for every DMU including DMU k, and the optimal value of $\mathbf{u}^{\mathrm{T}}\mathbf{y}_k/\mathbf{v}^{\mathrm{T}}\mathbf{x}_k$ is at most 1 (that is, 100%).

Since the virtual output-to-input ratio is not linear, the model (1) is a nonlinear programming problem. By imposing the constraint that the virtual input should be 1 (that is $\mathbf{v}^{\mathrm{T}}\mathbf{x}_k = 1$) and with an assumption that all inputs and outputs are positive, Charnes and Cooper (1962) suggested the model (1) be converted to the following linear programming model (2).

$$\left.\begin{array}{ll}\text{Maximize :} & \hat{w}_k = \mathbf{u}^{\mathrm{T}}\mathbf{y}_k \\ \text{Subject to :} & \\ & v^{\mathrm{T}}x_k = 1 \\ & \mathbf{u}^{\mathrm{T}}\mathbf{Y} - \mathbf{v}^{\mathrm{T}}\mathbf{X} \le \mathbf{0} \\ & \mathbf{u} \ge \varepsilon\mathbf{1} \\ & \mathbf{v} \ge \varepsilon\mathbf{1} \end{array}\right\} \qquad (2)$$

where

$\mathbf{Y} = (\mathbf{y}_1,\mathbf{y}_2,\ldots,\mathbf{y}_{\mathrm{n}})$: $(s \times n)$ matrix of observed outputs for all DMUs
$\mathbf{X} = (\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_{\mathrm{n}})$: $(m \times n)$ matrix of observed inputs for all DMUs
$\mathbf{0}$ = vector with all components equal to zero

In models (1) and (2), the only explicit restriction on the weights of **u** and **v** is positivity (i.e., $\mathbf{u} \ge \varepsilon\mathbf{1}$, $\mathbf{v} \ge \varepsilon\mathbf{1}$). This means that a priori specification of the weights is not required to evaluate efficiency. Rather, the solution of this model provides the best set of weights on **u** and **v** and the efficiency score $\hat{w}_k$ of DMU k. In order to obtain the optimal solutions of all DMUs, it is required to solve n models, each of which differs only in the coefficients of objective function. Thus the best set of weights and efficiency scores may vary from one to another DMU. These variable optimal weights do not allow the claims from DMUs that the evaluation process is unfair.

Model (2) can be converted into the following dual problem called the envelopment form using the linear programming duality theorem, as shown in model (3).

$$\left.\begin{array}{ll}\text{Minimize :} & \hat{\theta}_k = \theta - \varepsilon\mathbf{1}^{\mathrm{T}}\mathbf{s}^{+} - \varepsilon\mathbf{1}^{\mathrm{T}}\mathbf{s}^{-} \\ \text{Subject to :} & \\ & \mathbf{Y}\lambda - \mathbf{s}^{+} = \mathbf{y}_k \\ & \theta\mathbf{x}_k - \mathbf{X}\lambda - \mathbf{s}^{-} = \mathbf{0} \\ & \lambda, \mathbf{s}^{+}, \mathbf{s}^{-} \ge \mathbf{0} \end{array}\right\} \qquad (3)$$

where

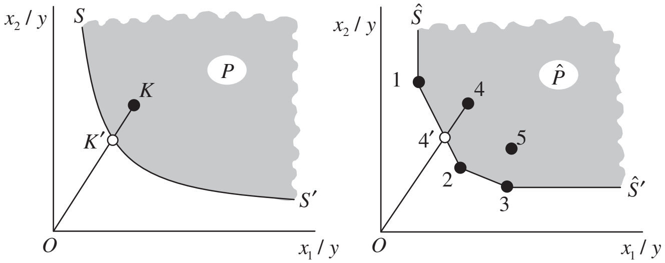
$\boldsymbol{\lambda}$ = $(n \times 1)$ vector of intensity variables for all DMUs
$\mathbf{s}^{+}$ = $(s \times 1)$ vector of output slack variables
$\mathbf{s}^{-}$ = $(m \times 1)$ vector of input slack variables

The value of $\hat{\theta}_k$ obtained will be the efficiency score for DMU k. The value is equal to $\hat{w}_k$ in the model (2) by the duality theorem. The slack variables, $\mathbf{s}^+$ and $\mathbf{s}^-$ mean output shortfalls and input excesses of DMUs. The presence of non-Archimedean ε in the objective function effectively allows the minimization over θ to preempt the optimization involving the slack variables. The solution of envelopment form is closely related to the definition of technical efficiency and its measure in the economics literature. If $\hat{\theta}_k = 1$, the DMU k is called technically efficient. Otherwise, the DMU k is called technically inefficient because a value of $\hat{\theta}_k < 1$ means that all inputs can be simultaneously reduced without altering the mix in which they are utilized.

In conventional economics, production technology is often described by the production possibilities set (PPS) which is a set of all input–output combinations that are technically feasible. Usually the PPS is assumed to satisfy convexity of itself and free disposability of inputs and outputs. A DMU operates either on the boundary of PPS if it is technically efficient, or inside the PPS if it is not technically efficient. An index of the technical efficiency introduced by Farrell (1957) is defined in terms of the distance between the DMU being evaluated and the projected point of the DMU into the boundary of PPS in condition of being given outputs. The boundary of PPS is not known in practice, and thus must be constructed from observed sample data. The envelopment form model (3) is a tool for constructing the boundary of PPS from the observed sample data and measuring the technical efficiency.

The Farrell index and the envelopment form of DEA can be illustrated with a simple example involving DMUs which use two inputs (x_1 and x_2) to produce a single output (y), under the assumption of a constant return to scale. The assumption of the constant return to scale allows the production technology to be represented using the unit iso-quant. The Farrell index assumes that the boundary of PPS, that is a set of technically efficient DMUs, is known. In panel (a) of Fig. 1, the curve SS' is the boundary of PPS P. For a given DMU k, the Farrell index of technical efficiency is defined as the ratio $\overline{OK'}/\overline{OK}$.

In practice, the boundary of PPS, SS' is not known. The envelopment form model (3) constructs the boundary of PPS from observations, and it also calculates the technical efficiency of a given DMU with reference to the constructed boundary of PPS. In panel (b) of Fig. 1, the piece-wise line $\hat{S}\hat{S}'$ is the constructed boundary of PPS $\hat{P}$ from the sample of observations 1, 2, 3, 4, and 5. The technical efficiency for DMU 4 is the ratio $\overline{O4'}/\overline{O4}$ which is equal to optimal value $\hat{\theta}_4$ of the model (3).



(a) Theoretical Measure of Technical Efficiency

(b) Empirical Measure of Technical Efficiency

Fig. 1. Measure of Technical Efficiency.

In Fig. 1, note that the constructed PPS is a subset of the unknown PPS, $\hat{P} \subseteq P$, and so the constructed boundary $\hat{S}\hat{S}'$ is an inward-biased estimator of the unknown boundary SS'. If the number of observations approaches to infinity, the constructed boundary $\hat{S}\hat{S}'$ approaches to the boundary SS'. This consistency property has been investigated by Korostelev, Simar, and Tsybakov (1995). For the DMU 4, $\hat{\theta}_4$ is an upward-biased estimator of the unknown technical efficiency value.

2.2. The Bootstrap

It is difficult to obtain analytically the sampling properties of the DEA efficiency index because of their complexity. In this case, the bootstrap introduced by Efron (1979) may be a useful tool. The bootstrap is a computationally intensive statistical inference method that can be used in many ways including establishment of bias and confidence intervals of complex estimator. Recently, much effort has been devoted to the bootstrap for DEA efficiency index (Atkinson & Wilson, 1995; Ferrier & Hirschberg, 1997; Lothgren & Tambour, 1999; Simar, 1996; Simar & Wilson, 1998). The idea of the bootstrap is to simulate repeatedly the data generating process using resampling method and applying the DEA to each simulated sample so that the resulting DEA efficiency indices mimic the sampling distribution of the original DEA efficiency indices. In this paper, we employ the bootstrap procedure for DEA efficiency index suggested by Simar and Wilson (1998).

The procedure for obtaining the bootstrap DEA efficiency indices $\hat{\theta}^*_{k,b}$ ($b = 1, 2,\ldots,B$, $k = 1, 2,\ldots,n$) from the original DEA efficiency indices $\hat{\theta}_k$ ($k = 1,2,\ldots,n$) is summarized in the appendix. The estimated bias $bias_k$ of the DEA efficiency index $\hat{\theta}_k$ for the kth DMU is

$$bias_k = \frac{1}{B}\sum_{b=1}^{B} \hat{\theta}^*_{k,b} - \hat{\theta}_k \tag{4}$$

where

$\hat{\theta}_k$ = the kth DMU's DEA efficiency index calculated from model (3)
$\hat{\theta}^*_{k,b}$ = the bth bootstrap DEA efficiency index for the kth DMU
B = the number of bootstrap replications

The bias-corrected DEA efficiency index for the kth DMU ($\tilde{\theta}_k$) is

$$\tilde{\theta}_k = \hat{\theta}_k - bias_k \tag{5}$$

After correction for bias, the bootstrap DEA efficiency indices $\hat{\theta}^*_{k,b}$ ($b = 1, 2,\dots,B$, $k = 1, 2,\dots,n$) can be used to provide the additional information on the uncertainty of rankings of railways in terms of operational efficiency. The correction for bias is obtained as follows:

$$\tilde{\theta}^*_{k,b} = \hat{\theta}^*_{k,b} - 2bias_k \tag{6}$$

The percentile method (Efron & Tibshirani, 1993) is the most straightforward method to obtain the bootstrap confidence intervals. The percentile method is based on the empirical cumulative distribution function $\hat{G}$ of $\tilde{\theta}^*_{k,b}$. The $1-2\alpha$ percentile confidence interval is given by

$$\left(\tilde{\theta}_k^{*(\alpha)}, \tilde{\theta}_k^{*(1-\alpha)}\right) \tag{7}$$

where $\tilde{\theta}_k^{*(\alpha)}$ indicates the α percentiles of $\hat{G}$.

The uncertainty information on the efficiency rankings of railways (Achieved Inversion Level) $AIL(i,j)$ can be obtained from the achieved inversion level between the ith DMU's bias-corrected DEA efficiency index $\tilde{\theta}_i$ and the jth DMU's bias-corrected DEA efficiency index $\tilde{\theta}_j$:

$$AIL(i,j) = \begin{cases} \#\{\tilde{\theta}^*_{i,b} - \tilde{\theta}^*_{j,b} < 0\}/B, & \tilde{\theta}_i - \tilde{\theta}_j > 0 \\ \#\{\tilde{\theta}^*_{i,b} - \tilde{\theta}^*_{j,b} \geq 0\}/B, & \tilde{\theta}_i - \tilde{\theta}_j < 0 \end{cases} \tag{8}$$

which is interpreted as the probability that the inversion occurs in all the bootstrap replications.

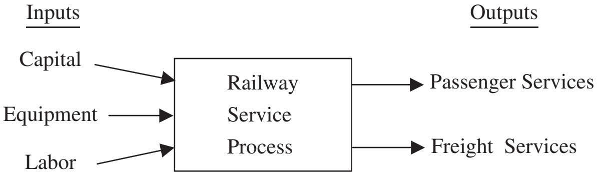


Fig. 2. Railway Service Process.

3. AN APPLICATION TO RAILWAY OPERATIONS

To evaluate the railway operational efficiency, we consider the railway operating systems in Fig. 2.

Capital is proxied using the length of lines, which is the sum of standard gauge as well as narrow and wide gauges and measured by kilometers. Equipment is represented by the rolling stock, that is the sum of passenger cars, cargo cars, and motor cars, which operating entities own and rent in the year. Labor is measured by the average number of persons who are directly paid by operating entities.

Our evaluation focuses on the country-based efficiency, not on company-based efficiency, because railway operating systems are different in each country. For example, Korean National Railroad (KNR) owns and operates railway lines, trains, and other equipment in Korea. However, in some countries like Sweden, railway lines and trains are owned and operated by different entities.

The data for this study (Table 1) is obtained from the UIC publications (Union Internationale des Chemins de fer, 2003) and includes 29 countries.

In Table 1, the passenger kilometers are generated by multiplying the number of paid passengers by the average distance per passenger. The cargo tonnage figures are generated by multiplying the weight of cargo by paid distance for operational efficiency comparisons.

4. MODEL RESULT

All computations were performed using MATLAB (The MathWorks Inc., 2000) code written by the authors and LINDO API (Lindo Systems Inc., 2001) as an optimization engine. The number of bootstrap replications B was 2000 and the required computational times were 5 min on a Pentium 4 PC.

Table 1. Input and Output Data for Railways Operations.

k	Country code	Country	Inputs			Outputs	
			Capital (length of lines, km)	Equipment (number of rolling stocks)	Labor (number of persons)	Passenger services (million passenger-km)	Freight services (million ton-km)
1	AT	Austria	5,643	29,648	52,272	8,080	15,937
2	BE	Belgium	3,472	24,877	40,606	7,354	8,569
3	CH	Switzerland	2,902	25,638	30,827	13,136	10,154
4	CZ	Czech	9,365	73,953	89,220	6,929	16,464
5	DE	Germany	37,525	221,414	280,044	72,822	71,593
6	DK	Denmark	2,324	3,397	10,470	5,113	1,938
7	ES	Spain	12,319	35,910	38,604	19,245	13,492
8	FI	Finland	5,836	14,388	13,548	3,415	9,778
9	FR	France	31,589	142,014	174,447	66,298	55,863
10	GR	Greece	2,299	4,176	10,523	1,583	347
11	HR	Croatia	2,726	11,578	19,468	943	1,849
12	HU	Hungary	7,768	29,694	58,048	6,835	7,909
13	IE	Ireland	1,919	2,473	5,234	1,421	526
14	IT	Italy	16,108	92,821	114,193	40,971	25,450
15	JP	Japan	20,165	68,728	172,203	241,733	22,313
16	KR	Korea	3,098	20,021	31,960	28,356	10,072
17	LU	Luxembourg	274	2,685	3,050	300	623
18	MY	Malaysia	1,622	4,587	6,000	6,300	907
19	NL	Netherlands	2,808	8,418	26,286	14,330	3,549
20	NO	Norway	4,179	4,202	10,444	2,674	2,862
21	PL	Poland	22,891	114,033	203,987	21,518	55,482
22	PT	Portugal	2,813	6,194	12,816	4,329	2,562
23	RO	Romania	11,364	155,340	105,460	12,304	14,680
24	SE	Sweden	9,978	21,638	25,187	7,434	18,490
25	SI	Slovenia	1,202	7,287	9,037	623	2,784
26	SK	Slovakia	3,662	34,169	48,913	2,968	9,914
27	TR	Turkey	8,682	20,542	42,721	6,146	8,446
28	TW	Taiwan	1,104	8,014	16,419	9,978	1,279
29	UA	Ukraine	22,473	244,015	367,878	47,600	156,336

Table 2 presents the results of DEA applications and the bootstrap, which includes the DEA efficiency index $\hat{\theta}_k$ (column 4), the corresponding bias estimates $bias_k$ (column 5), and the bias-corrected DEA efficiency index $\tilde{\theta}_k$ (column 6). The last two columns (columns 7 and 8) exhibit the information on the estimated 95% confidence interval $(\tilde{\theta}_k^{*(\alpha)}, \tilde{\theta}_k^{*(1-\alpha)})$.

Table 2. Results of DEA and the Bootstrap.

k	Country Code	Country	DEA Efficiency Index$\hat{\theta}_k$	Bias $bias_k$	Bias-corrected DEA efficiency index $\tilde{\theta}_k$	Percentile Confidence Interval	
						2.5% $\tilde{\theta}_k^{*(\alpha)}$	97.5% $\tilde{\theta}_k^{*(1-\alpha)}$
1	AT	Austria	0.7874	0.0667	0.7207	0.6560	0.8645
2	BE	Belgium	0.5510	0.0795	0.4715	0.3939	0.6076
3	CH	Switzerland	0.8512	0.1213	0.7300	0.6118	0.9537
4	CZ	Czech	0.4034	0.0417	0.3617	0.3213	0.4469
5	DE	Germany	0.5835	0.0516	0.5319	0.4825	0.6355
6	DK	Denmark	0.9012	0.0857	0.8155	0.7326	1.0280
7	ES	Spain	0.6942	0.1146	0.5796	0.4670	0.7558
8	FI	Finland	0.9831	0.2638	0.7193	0.4575	1.1940
9	FR	France	0.6880	0.0782	0.6097	0.5336	0.7351
10	GR	Greece	0.1605	0.0327	0.1278	0.0955	0.2005
11	HR	Croatia	0.2252	0.0204	0.2048	0.1852	0.2427
12	HU	Hungary	0.3838	0.0282	0.3556	0.3285	0.4152
13	IE	Ireland	0.3385	0.0484	0.2901	0.2426	0.3868
14	IT	Italy	0.5498	0.0689	0.4809	0.4140	0.6024
15	JP	Japan	1.0000	0.6046	0.3954	-0.2062	1.3740
16	KR	Korea	1.0000	0.3014	0.6986	0.4006	1.2140
17	LU	Luxembourg	0.4659	0.0555	0.4104	0.3562	0.5439
18	MY	Malaysia	0.8086	0.2228	0.5858	0.3657	1.0671
19	NL	Netherlands	0.7916	0.1615	0.6301	0.4713	0.9499
20	NO	Norway	0.8634	0.0978	0.7656	0.6702	0.9451
21	PL	Poland	0.6958	0.0623	0.6335	0.5735	0.7649
22	PT	Portugal	0.5816	0.0755	0.5061	0.4323	0.6203
23	RO	Romania	0.3251	0.0255	0.2996	0.2752	0.3610
24	SE	Sweden	1.0000	0.3415	0.6585	0.3203	1.0925
25	SI	Slovenia	0.6270	0.0704	0.5566	0.4880	0.6694
26	SK	Slovakia	0.4631	0.0987	0.3644	0.2672	0.5648
27	TR	Turkey	0.5149	0.0694	0.4455	0.3780	0.5561
28	TW	Taiwan	0.7858	0.2205	0.5653	0.3474	1.1464
29	UA	Ukraine	1.0000	0.3542	0.6458	0.2948	1.2224
		Average	0.6560	0.1332	0.5228		

The average of the 29 DEA efficiency indices $\hat{\theta}_k$ (column 4) is 0.6560. Of 29 countries, four railway systems including JP, KR, SE, and UA appear ostensibly efficient, as indicated by DEA efficiency indices $\hat{\theta}_k = 1$. The remaining 25 railway systems appear relatively inefficient with different degrees of magnitude, as indicated by DEA efficiency indices ranging from

0.1605 to 0.9831. As noted in Section 2.1, the DEA efficiency index is upward-biased. So ignoring the statistical property can lead to erroneous conclusions. The bias estimates $bias_k$ (column 5) in Table 2 reveal the sensitivity of the DEA efficiency index with respect to the sampling variations. The average of 29 bias estimates $bias_k$ (column 5) is 0.1332. This indicates that the DEA efficiency index overestimates the operational efficiency of railways.

The average of the 29 bias-corrected DEA efficiency indices $\tilde{\theta}_k$ (column 6) is 0.5528. The railways whose DEA efficiency indices are unity, appear to be inefficient, as indicated by the bias-corrected DEA efficiency index ranging from 0.3954 to 0.6986. For example, KR (Korea) appears ostensibly efficient if only the DEA efficiency index is examined. However, KR has a bias-corrected DEA efficiency index of 0.6986 suggesting that the same outputs (passenger and cargo service) could have been produced, while the scaling inputs (lines, car, and manpower) are adjusted by more than 30.14% $[=(1-0.6986)\times 100]$. On the other hand, DK (Denmark) whose DEA efficiency index is not unity appears to be best operated, as indicated by the ranking of the bias-corrected DEA efficiency index.

Fig. 3 illustrates a graphical representation of the distribution of $\tilde{\theta}^*_{k,b}$ ($b = 1,2,...,2000$, $k = 1,2...,29$) which shows box plots to facilitate the comparison among the 29 railway systems' operational efficiencies.

For each railway system in Fig. 3, the size of box is determined by the range of values 25th–75th percentiles. The size of boxes varies quite somewhat across the 29 railway systems. This indicates how sensitive a particular railway system's DEA efficiency index is to sampling variations. For example, the box sizes of Austria, Belgium, and Switzerland are relatively smaller than those of Finland, Japan, and Korea, implying that the DEA efficiency indices of Austria, Belgium, and Switzerland are more precise than those of others.

In Table 2, Taiwan has a DEA index score of 0.7858 while Ukraine is efficient with a DEA efficiency score of 1.0000. However, the last two columns of these two countries show that the percentile confidence intervals for the operational efficiency of the two railway systems overlap to a large degree. Thus, we would not say that the two railway systems are significantly different in terms of their operational efficiency. Table 3 provides information on the uncertainty of the efficiency rankings of railway systems that can be obtained from the achieved inversion level $AIL(i,j)$. This can be interpreted as the probability that the efficiency rankings inversion occurs in all the bootstrap replications. For example, Denmark has a bias-corrected

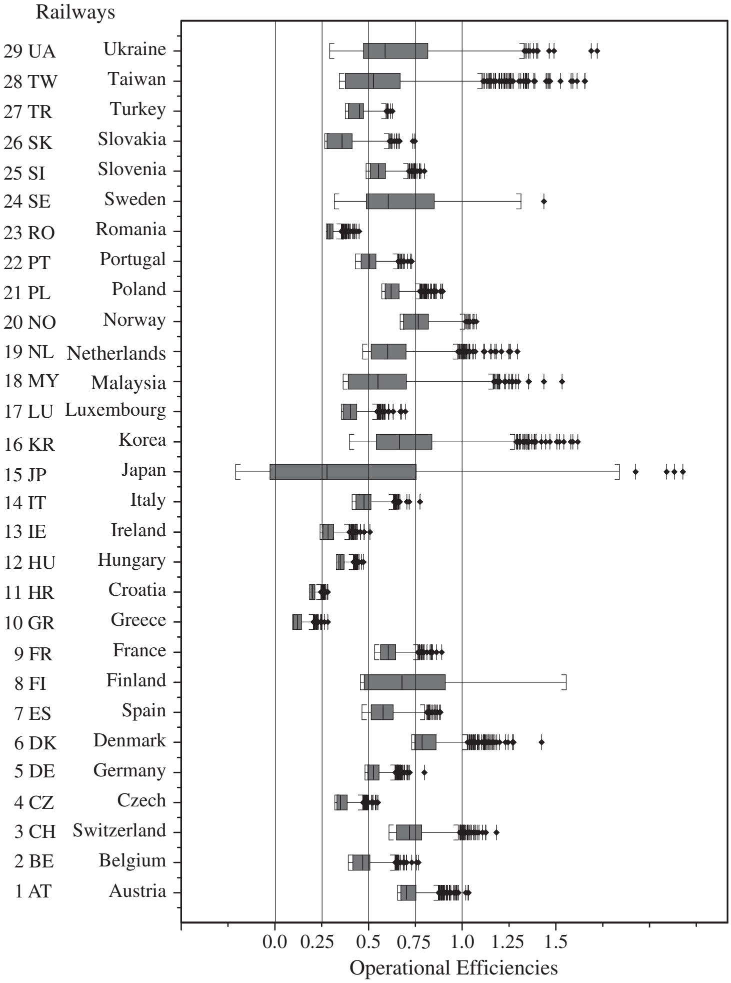


Fig. 3. Box Plots of Railway Operational Efficiencies.

DEA efficiency index $\tilde{\theta}_1 = 0.8155$, while Norway has a bias-corrected DEA efficiency index $\tilde{\theta}_2 = 0.7656$. The achieved inversion level $AIL(1, 2)$ between Denmark ($k = 1$) and Norway ($k = 2$) is 0.30. This implies that the probability that the efficiency rankings inversion occurs in all the bootstrap replications.

Table 3. Achieved Inversion Level.

k	Country Code	Country	Bias-corrected DEA Efficiency Index $\tilde{\theta}_k$	Achieved Inversion Level						
				AIL $(k,k+1)$	*AIL* $(k,k+2)$	*AIL* $(k,k+3)$	*AIL* $(k,k+4)$	*AIL* $(k,k+5)$	*AIL* $(k,k+6)$	*AIL* $(k,k+7)$
1	DK	Denmark	0.8155	0.30	0.23	0.15	0.31	0.27	0.29	0.26
2	NO	Norway	0.7656	0.37	0.29	0.37	0.36	0.31	0.33	0.07*
3	CH	Switzerland	0.7300	0.49	0.43	0.38	0.39	0.37	0.14	0.22
4	AT	Austria	0.7207	0.43	0.41	0.41	0.36	0.00**	0.24	0.08*
5	FI	Finland	0.7193	0.48	0.30	0.42	0.44	0.43	0.41	0.33
6	KR	Korea	0.6986	0.45	0.42	0.42	0.33	0.37	0.28	0.32
7	SE	Sweden	0.6585	0.45	0.51	0.47	0.46	0.42	0.39	0.38
8	UA	Ukraine	0.6458	0.35	0.50	0.51	0.44	0.46	0.42	0.43
9	PL	Poland	0.6335	0.43	0.38	0.35	0.28	0.26	0.10*	0.04**
10	NL	Netherlands	0.6301	0.50	0.34	0.38	0.28	0.37	0.29	0.12
11	FR	France	0.6097	0.39	0.38	0.33	0.21	0.02**	0.06	0.00**
12	MY	Malaysia	0.5858	0.57	0.36	0.52	0.48	0.43	0.38	0.37
13	ES	Spain	0.5796	0.37	0.40	0.32	0.19	0.12	0.15	0.05**
14	TW	Taiwan	0.5653	0.60	0.52	0.47	0.42	0.41	0.36	0.30
15	SI	Slovenia	0.5566	0.29	0.24	0.13	0.08*	0.04**	0.01**	0.38
16	DE	Germany	0.5319	0.34	0.07*	0.15	0.07*	0.02**	0.40	0.03**
17	PT	Portugal	0.5061	0.34	0.31	0.12	0.09*	0.41	0.08*	0.01**
18	IT	Italy	0.4809	0.43	0.31	0.13	0.42	0.10*	0.02**	0.00**
19	BE	Belgium	0.4715	0.36	0.10*	0.43	0.08	0.00**	0.00**	0.00**
20	TR	Turkey	0.4455	0.27	0.45	0.19	0.09*	0.01**	0.00**	0.01**
21	LU	Luxembourg	0.4104	0.47	0.24	0.00**	0.11	0.00**	0.11	0.00**
22	JP	Japan	0.3954	0.52	0.52	0.52	0.51	0.50	0.43	0.40
23	SK	Slovakia	0.3644	0.55	0.51	0.30	0.23	0.00**	0.00**	
24	CZ	Czech	0.3617	0.51	0.00**	0.08*	0.00**	0.00**		
25	HU	Hungary	0.3556	0.04**	0.06*	0.00**	0.00**			
26	RO	Romania	0.2996	0.39	0.00**	0.00**				
27	IE	Ireland	0.2901	0.00**	0.00**					
28	HR	Croatia	0.2048	0.02**						
29	GR	Greece	0.1278							

*Indicates that *AIL*(i,j) is between 0.05 and 0.10.

**Indicates that *AIL*(i,j) is between 0.00 and 0.05.

5. CONCLUSION

We applied both DEA and the bootstrap analyses to railway operation data of 29 UIC countries, which provide input and output information required for this study. The average of the 29 DEA efficiency index is 0.6560, while the average of 29 bias estimates is 0.1332. This implies that the DEA efficiency index overestimates the operational efficiency of railway systems under study. The railway systems whose DEA efficiency indices are unity, appear to be inefficient in terms of bias-corrected DEA efficiency index. However, some railway systems such as DK (Denmark) whose DEA efficiency index is not unity appears to be relatively well operated in terms of the ranking of the bias-corrected DEA efficiency index. In this paper, only 29 UIC countries are evaluated as part of social overhead capital efficiency. Therefore, the result of this paper is only valid for the relative efficiency measures among the 29 UIC countries. If other countries are included, the operational efficiency may be changed. To evaluate the ability of railway operational managers, the time series data should be used for every country and compare the improvement by each year.

REFERENCES

Atkinson, S. E., & Wilson, P. W. (1995). Comparing mean efficiency and productivity scores from small samples: A bootstrap methodology. *Journal of Productivity Analysis*, *6*(2), 137–152.

Charnes, A., & Cooper, W. W. (1962). Programming with linear fractional functionals. *Naval Research Logistics Quarterly*, *9*(3/4), 181–185.

Charnes, A., Cooper, W. W., & Rhodes, E. L. (1978). Measuring the efficiency of decision making units. *European Journal of Operational Research*, *2*(6), 429–444.

Cooper, W. W., Huang, Z., Lelas, V., Li, S. X., & Olesen, O. B. (1998). Chance constrained programming formulations for stochastic characterizations of efficiency and dominance in DEA. *Journal of Productivity Analysis*, *9*(1), 53–79.

Efron, B. (1979). Bootstrap methods: Another look at the Jackknife. *Annals of Statistics*, *7*(1), 1–26.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap,* Monographs on Statistics and Applied Probability, No. 57. New York: Chapman & Hall.

Elkins, T. Y., & Lawrence, K. D. (2001). Measuring performance: A multiple criteria/multiple objective data envelopment analysis approach. In: K. D. Lawrence, G. R. Reeves & J. B. Guerard, Jr. (Eds), *Advances in Mathematical Programming and Financial Planning* (Vol. 6, pp. 35–53). Netherlands: Elsevier Science/JAI Press, Amsterdam.

Farrell, M. J. (1957). The measurement of productive efficiency. *Journal of the Royal Statistical Society, Series A, General*, *120*(3), 253–281.

Ferrier, G. D., & Hirschberg, J. G. (1997). Bootstrapping confidence intervals for linear programming efficiency scores: With an illustration using italian banking data. *Journal of Productivity Analysis*, *8*(1), 19–33.

Gstach, D. (1995). Comparing structural efficiency of unbalanced subsamples: A resampling adaptation of data envelopment analysis. *Empirical Economics*, *20*(3), 531–542.

Holvad, T., Hougard, J. L., Kronborg, D., & Kvist, H. K. (2004). Measuring inefficiency in the Norwegian bus industry using multidirectional efficiency analysis. *Transportation*, *31*(3), 349–369.

Karlaftis, M. G. (2004). A DEA approach for evaluating efficiency and effectiveness of urban transit system. *European Journal of Operational Research*, *152*, 354–364.

Korostelev, A., Simar, L., & Tsybakov, A. (1995). On estimation of monotone and convex boundaries. *Publications de l'Institute de statistique de l'Universite de Paris*, *39*, 3–18.

Kwak, N. K., & Kim, S. H. (2000). Data envelopment analysis: Concepts, applications, and perspectives. In: S. B. Dahiya (Ed.), *The current state of business disciplines*, (Vol. 2, pp. 519–535). Mohtak, India: Spellbound Publications, Ltd.

Kwak, N. K., Choi, T. S., & Kim, S. H. (2001). Efficiency evaluation of research university libraries using data envelopment analysis. In: K. D. Lawrence, G. R. Reeves & J. B. Guerard, Jr. (Eds.), *Advances in Mathematical Programming and Financial Planning*, (Vol. 6, pp. 3–18). The Netherlands: Elsevier Science/JAI Press, Amsterdam.

Land, K. C., Lovell, C. A. K., & Thore, S. (1993). Chance-constrained data envelopment analysis. *Managerial and Decision Economics*, *14*(6), 541–554.

Li, S. X. (1998). Stochastic models and variable returns to scales in data envelopment analysis. *European Journal of Operational Research*, *104*(3), 532–548.

Lindo Systems Inc. (2001). *LINDO callable library: The premier optimization engine*. Chicago, IL: Lindo Systems Inc.

Lothgren, M., & Tambour, M. (1999). Bootstrapping the data envelopment analysis malmquist productivity index. *Applied Economics*, *31*(4), 417–425.

Martin, J. C., Gutierrez, J., & Roman, C. (2004). Data envelopment analysis (DEA) index to measure the accessibility impacts of new infrastructure investments: The case of high-speed train corridor Madrid-Barcelona-French Border. *Regional Studies*, *38*(6), 697–723.

Olesen, O. B., & Petersen, N. C. (1995). Chance constrained efficiency evaluation. *Management Science*, *41*(3), 442–457.

Seiford, L. M. (1996). Data envelopment analysis: The evolution of the state of the art (1978–1995). *Journal of Productivity Analysis*, *7*(2–3), 99–137.

Seiford, L. M., & Zhu, J. (1998). An acceptance system decision rule with data envelopment analysis. *Computers & Operations Research*, *25*(4), 329–332.

Simar, L. (1996). Aspects of statistical analysis in DEA-type frontier models. *Journal of Productivity Analysis*, *7*(2/3), 177–185.

Simar, L., & Wilson, P. W. (1998). Sensitivity analysis of efficiency scores: How to bootstrap in nonparametric frontier models. *Management Science*, *44*(1), 49–61.

The Math Works Inc. (2000). *Using MATLAB*. Natick, MA.: The MathWorks Inc.

Troutt, M. D., Rai, A., & Zhang, A. (1996). The potential use of DEA for credit applicant acceptance systems. *Computers & Operations Research*, *23*, 405–408.

Union Internationale des Chemins de fer (2003). *International Railway Statistics*, (Statistics of Individual Railways), Paris, France.

APPENDIX

Simar and Wilson (1998) suggested a bootstrap procedure for analyzing the sampling properties of DEA efficiency index, which is based on the smoothing and reflection method in the resampling. The bootstrap procedure can be summarized as follows:

(1) For each railway $(\mathbf{x}_k, \mathbf{y}_k)$ of the original data calculate the operational efficiency index $\hat{\theta}_k$ by the DEA, and transform the input–output vectors $(\mathbf{x}_k, \mathbf{y}_k)$, $k = 1, 2, ..., n$ using the DEA efficiency indices $\hat{\theta}_k$, $k = 1, 2, ..., n$ as $(\hat{\mathbf{x}}_k^f, \mathbf{y}_k) = (\hat{\theta}_k \cdot \mathbf{x}_k, \mathbf{y}_k)$.

(2) Initialize the bootstrap replication index b as $b = 0$.

(3) Increment the value of bootstrap replication index b as $b = b + 1$.

(4)Generate a random sample of size n, $\theta^*_{1,b}, \theta^*_{2,b}, ..., \theta^*_{n,b}$, from $\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_n$ as follows:

(4a) Given the set of calculated DEA efficiency indices $\hat{\theta}_1$, $\hat{\theta}_2, ..., \hat{\theta}_n$, obtain the bandwidth parameter h such that

$$h = 0.9n^{-1/5} \min\{\hat{\sigma}_{\hat{\theta}}, R_{13}/1.34\}$$

where $\hat{\sigma}_{\hat{\theta}}$ denotes the plug-in standard deviation of the DEA efficiency indices $\hat{\theta}_1$, $\hat{\theta}_2, ..., \hat{\theta}_n$, and R_{13} denotes the interquartile range of the empirical distribution of $\hat{\theta}_1$, $\hat{\theta}_2, ..., \hat{\theta}_n$.

(4b) Generate $\beta^*_{1,b}, \beta^*_{2,b}, ..., \beta^*_{n,b}$ by resampling with replacement from the empirical distribution of $\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_n$.

(4c) Define the sequence $\tilde{\theta}^*_1, \tilde{\theta}^*_2, ..., \tilde{\theta}^*_n$ using the following:

$$\tilde{\theta}^*_j = \begin{cases} \beta^*_j, & \text{if } \beta^*_j + h\varepsilon^*_j \le 1 \\ 2 - \beta^*_j - h\varepsilon^*_j, & \text{otherwise} \end{cases}, \quad j = 1, 2, ..., n$$

(4d) Define the bootstrap sequence $\theta^*_{1,b}, \theta^*_{2,b}, \ldots, \theta^*_{n,b}$ using the following:

$$\theta^*_{j,b} = \bar{\beta}^*_b + \frac{1}{\sqrt{1 + h^2 \big/ \hat{\sigma}^2_{\hat{\theta}}}} \left(\tilde{\theta}^*_j - \bar{\beta}_b \right), \quad j = 1, 2, ..., n$$

where $\bar{\beta}^*_b = (1/n) \sum_{j=1}^{n} \beta^*_{j,b}$.

(5) Compute the bootstrap pseudo-sample data $(\mathbf{x}^*_{j,b}, \mathbf{y}_j)$, $j = 1, 2, ..., n$, using the following:

$$(\mathbf{x}^*_{j,b}, \mathbf{y}_j) = \left(\frac{1}{\theta^*_{j,b}} \cdot \mathbf{x}^f_j, \mathbf{y}_j \right)$$

(6) Compute the bootstrap DEA efficiency index $\hat{\theta}^*_{k,b}$ for $k = 1, 2, \ldots, n$ using the pseudo-sample data and the following linear program:

$$\begin{aligned} &\text{Minimize :} && \theta^*_{k,b} = \theta - \varepsilon \mathbf{1}^{\mathrm{T}} \mathbf{s}^+ - \varepsilon \mathbf{1}^{\mathrm{T}} \mathbf{s}^- \\ &\text{Subject to :} && \\ & && \mathbf{Y}\lambda - \mathbf{s}^+ = \mathbf{y}_k \\ & && \theta \mathbf{x}_k - \mathbf{X}^*_b \lambda - \mathbf{s}^- = \mathbf{0} \\ & && \lambda, s^+, s^- \ge 0 \end{aligned}$$

where

$\mathbf{Y} = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n)$: $(s \times n)$ matrix of observed output data
$\mathbf{X}^*_b = (\mathbf{x}^*_{1,b}, \mathbf{x}^*_{2,b}, \ldots, \mathbf{x}^*_{n,b})$: $(m \times n)$ matrix of the bootstrap input data

λ:($n \times 1$) vector of intensity variables for all DMUs
$\mathbf{s}^{+}$:($s \times 1$) vector of output slack variables
$\mathbf{s}^{-}$:($m \times 1$) vector of input slack variables

(7) Until $b = B$ repeat steps (3)–(6) to provide for $k = 1, 2, \ldots, n$ a set of bootstrapped efficiency indices $\{\hat{\theta}^{*}_{k,b}, b = 1, 2, ..., B\}$.

INPUT–OUTPUT VARIABLE EVALUATION FOR STUDYING DEA EFFICIENCY PERFORMANCE IN ELECTRIC UTILITIES

David E. Schultz and Suresh K. Tadisina

ABSTRACT

Researchers and corporate managers are interested in firm performance and its measurement. Data envelopment analysis (DEA) offers a powerful analytical tool that can be utilized to identify and measure firm performance. This paper provides a comprehensive set of key input and output variables necessary for DEA analysis of electric utility performance.

INTRODUCTION

In this day and time many researchers and corporate managers are interested in firm performance and its measurement. With the recent deregulation and move to competition this is no more apparent now than ever with the electric utility industry and within the individual electric utility companies. With the collapse and failure of ENRON, this interest will no doubt heighten and become more intense. Data envelopment analysis (DEA) offers a valuable and powerful analytical tool that can be utilized to identify and

Applications of Management Science: In Productivity, Finance, and Operations
Applications of Management Science, Volume 12, 41–47

ISSN: 0276-8976/doi:10.1016/S0276-8976(06)12004-0

measure electric utility performance. This analytical tool requires as a basic input, the identification and quantification of a set of key critical input and output variables for the individual entities to be studied. This paper provides a set of key input and output variables necessary to provide for the analysis of electric utility firm performance.

Cooper, Seiford, and Tone (2000) have prepared a detailed bibliography of research studies and publications that include over 1,500 different analyses utilizing DEA and other methods of evaluating firm performance. A review of the titles of the various publications revealed that there is much interest in the behavior of organizations and industries with respect to performance. Efficiency was the most cited issue accounting for approximately 61 percent of those reviewed. Performance and productivity were also found to be of major interest accounting for approximately 30 percent of the publications.

PERFORMANCE

High productivity is critical also for the prosperity and well-being of an organization or firm in its own industry and markets. Organizations and firms competing in this fast-paced changing international environment must continue to find ways to be competitive in order to survive and prosper. There are three primary needs that an organization must address to succeed. These are the need for improved productivity, the need for improved flexibility, and the need to develop competitive advantages (Lee & Schneiderjans, 1993, pp. 5–6).

Bitran and Chang (1984) view performance of the firm as an input–output conversion process and consider productivity and efficiency of production as indicators and measures of firm performance. Much of the literature follows this view (Elion, 1985; Markland, Vickery, & Davis, 1998; Troutt, Rai, Tadisina, & Zhang, 1998; Van Zandlt, 1998). Productivity and efficiency of the firm are to be measured by consideration of a ratio measure of outputs and inputs. Researchers differ in the literature as to the consideration of key firm inputs and outputs as well as the weights assigned to those respective variables.

PERFORMANCE MEASUREMENT

Evolution of DEA

Farrell (1957) utilized this economic production function methodology to study farm production by assessing and comparing agricultural production in each of the 48 states in the United States. His model utilized one output (agricultural production) and four inputs (land, labor, materials, and capital). He extended the shape of the production function frontier to the mathematical Cobb–Douglas model and broadened the application to include multiple outputs and multiple inputs.

In an effort to ease the solution of such ratio problems encountered in firm performance analysis, Charnes and Cooper (1962) developed a methodology to convert the fractional linear programming problem into two separate linear programming (LP) formulations. One such LP model would be representative of the numerator (or output(s)) and the other LP model would be representative of the denominator (or input(s)). As long as the function could be treated as consisting of linear piecewise segments, the two respective LP models could be evaluated either by maximizing the numerator in order to optimize the ratio or by minimizing the denominator in order to maximize the ratio or efficiency under consideration. Either maximizing the efficiency so as to provide a maximum output with the given input resources or attaining the same output level with a minimum of input resources results in maximizing the overall efficiency of the given firm or group of firms described by the linear fractional programming problem.

In extending the work of Farrell (1957) and the application of mathematical programming, Aigner and Chu (1968) developed a production function using the various factors of production for both inputs and outputs for the various firms under consideration. Their central focus provided a way to define an industry production function that would provide a relative measure of efficiency for the industry as well as for each firm utilizing the existing state-of-the-art technology available for the short term. The industry production function or frontier provided the basis for measuring and evaluating industry relative efficiency and firm relative efficiency in producing maximum output attainable with available technology.

Most of the early work in the pursuit of economic production functions and efficiency measurement was undertaken with known definitive input and output data values from previous historical periods. This historical data was treated as deterministic data in these analyses. Afriat (1972) suggested that another approach using stochastic analysis be incorporated in this work. Stochastic analysis could be applied to production data using probability and probability density functions for estimation of production functions and efficiency estimation and measurement.

The four most popular nonparametric DEA models according to the literature are the CCR ratio model (Charnes, Cooper, & Rhodes, 1978), the BCC ratio model (Banker, Charnes, & Cooper, 1984), the additive model, and the multiplicative models. Each of these models uses a nonparametric approach in estimating the efficiencies of one or more Decision Making Units (DMUs) in the sample or group under study. "In summary, the choice of a particular DEA model determines (1) the implicit returns-to-scale properties; (2) the geometry of the envelopment surface (with respect to which efficiency measurements will be made); and (3) the efficient projection, i.e., the inefficient DMU's path to the efficient frontier" (Charnes, Cooper, Lewin, & Seiford, 1994, pp. 45–46).

Prior Utility Studies

Several previous studies have been conducted with respect to evaluating utility performance. These studies have utilized historical firm specific input and output variables in various DEA models and statistical regression models. These studies considered various key input variables for the respective organizations utilized in their respective samples. However, the input variables selected did not represent a complete selection of the critical inputs utilized by the firm required to explain the transformation process of the firm. The input variables for the proposed study account for the total costs of the firm. The expense components necessary to track and explain both the fixed and variable expenses of the firm on an annual basis are tracked. Such treatment enables the researcher to assess fixed costs as well as variable costs of production and operation. Key system characteristics also are included in the proposed study that allow the researcher to evaluate and assess the size and investment capacity of the respective firms with respect to both its generation mix and transmission delivery system. It is important also to note that the database collection efforts in the proposed study will enable the researcher to make comparisons of various sizes and types of generating capacity. The type and sizes of various capacity, that is, coal-, gas-, and oil-fired, pumped storage, hydro, and/or nuclear generation capability and the requisite additions and/or retirements can be monitored and tracked on an annual basis.

The key output variables considered by prior studies are not adequate to account for the key outputs for electric utilities. The proposed study considers total kilowatthours of energy sold, maximum kilowatt system demand, total electric revenue, and net generation in kilowatthours for each respective firm on an annual basis. Notice that use of the total revenue as a key output with the total expenses as inputs enables the researcher to determine net margins and various component unit costs and revenues or rates among the sample sets. Such measures and capability provide additional checks within the model framework on the study results. The database collection also will gather information on maximum electric system demands that will allow the researcher the opportunity to determine when the maximum demands on the individual systems occur. The ability to determine whether a firm is winter-peaking or summer-peaking can provide additional insights to the nature of the electric systems under study and review. In a similar manner it also will be possible to determine the annual system load factor based on the system peak demand experienced by each firm. The load factor can provide a measure of how intensive the firm's electric load is utilizing its own system and infrastructure.

Furthermore, the firms included in the proposed sample consist of both the electric production and transmission delivery system side of the business for investor-owned utilities and rural electric G&T firms. The other studies are concerned primarily with the study and evaluation of organizations of like kind. While this is commendable for research, the inclusion of mixed organizations in the sample serves to provide representative sample organizations that a firm will likely encounter as its competitors in its marketplace under open competition and deregulation. Similar findings can be observed in considering the various output variables in each of the respective studies.

In addition, the proposed study is more pervasive and inclusive than prior studies. The proposed study will perform detailed DEA cross-sectional analysis for each period in the 1988 through 1997 horizon. Additionally, longitudinal DEA windows analyses using a three-year moving average window will be performed throughout the horizon to determine performance trends within the sample. These trends will assess performance for each utility separately as well as for the IOUs', the G&T's, and the sample as a whole. These other studies do not provide as robust analyses as the one contemplated in this proposal. An assessment also will be undertaken considering large as well as small size firms throughout this horizon. In order to provide reliability and validity checks on these DEA analyses, it is further proposed that a separate maximal decision efficiency model, the MER model, be implemented to test and compare with the DEA CCR model results. The other studies did not utilize this type of verification and validation analyses in their approaches. The proposed study also intends to evaluate the best in class and the inefficient firms to determine rationales for the conduct of the various firm performances. The hope is that insights can be obtained to help define and explain firm performance. Such information and insights would prove valuable for management and the organization for continuous improvement in performance to insure its success and survival.

The prior utility studies are each discussed and reviewed and are available upon request.

FUTURE RESEARCH PROPOSAL

It is recommended that a research study be conducted examining performance of electric utilities within the electric utility industry as suggested in this paper. Performance in such a research proposal would be concerned primarily with development a relative efficiency or productivity ratio for each firm using a set of selected inputs and outputs critical to the operation and success of the firm. It is also proposed that the specific nine input and four output variables identified be utilized simultaneously in this analysis. The ratio of the outputs to the inputs for a particular firm provides a measure of the efficiency of the organization. Comparisons of various firms' efficiency ratios during a single common time period provide a means of evaluating their relative efficiencies. Comparing such ratios in this single time period allows the researcher an opportunity to determine the most efficient frontier and to determine the best performing organizations and the poorest ones. This type of analysis is commonly referred to as a static or cross-sectional analysis.

Such a study will enable the researcher to determine the relative efficient firms that lie on the efficient frontier. This result also will identify the changes that the inefficient firms need to consider in order to move into the efficient frontier. In this way, the management of the firm can develop added insights into these critical areas that warrant additional attention and make such improvements.

Enlarging the time horizon for the examination of the various firms' relative efficiencies over a longer time period can expand this study effort. Such an investigation provides the opportunity to compare relative efficiencies of a single firm over time. If there are events or market changes that occur at discrete points in time, then organizational response to these events or changes can be assessed. Analysis of the relative efficient frontiers also can be examined over time. It is conceivable that the efficient frontier may be changing over time, and this type of analysis provides a way to examine such changes. This type of analysis is commonly referred to as a longitudinal analysis.

Norton (1994) proposed that the relative efficiency approach be undertaken to study the performance of individual functions within a single firm and/or individual subunit performance within a single firm. Troutt et al.

(1998) suggested a similar approach for evaluating investment in information technology in firms in competition with one another cross-sectionally in a single time period and/or longitudinally.

REFERENCES

Afriat, S. N. (1972). Efficiency estimation of production functions. *International Economic Review*, *13*(3), 568–598.

Aigner, D. J., & Chu, S. F. (1968). On estimating the industry production function. *The American Economic Review*, *58*(4), 826–839.

Banker, R. D., Charnes, A., & Cooper, W. W. (1984). Some models for estimating technical and scale inefficiencies in data envelopment analysis. *Management Science*, *30*(9), 1078–1093.

Bitran, G. R., & Chang, U. (1984). Productivity measurement at the firm level. *Interfaces*, *14*(3), 29–40.

Charnes, A., & Cooper, W. W. (1962). Programming with linear fractional functionals. *Naval Research Logistics Quarterly*, *9*(3), 181–186.

Charnes, A., Cooper, W. W., Lewin, A. Y., & Seiford, L. M. (Eds) (1994). *Data envelopment analysis: Theory, methodology, and application*. Boston: Kluwer Academic Publishers.

Charnes, A., Cooper, W. W., & Rhodes, E. (1978). Measuring the efficiency of decision making units. *European Journal of Operational Research*, *2*, 429–444.

Cooper, W. W., Seiford, L. M., & Tone, K. (2000). *Data envelopment analysis: A Comprehensive text with models, applications, references, and DEA-Solver software*. Boston: Kluwer Academic Publishing.

Eilon, S. (1985). A framework for profitability and productivity. *Interfaces*, *15*(3), 31–40.

Farrell, M. J. (1957). The measurement of productive efficiency. *Journal of the Royal Statistical Society*, *120*, 253–281.

Lee, S. M., & Schneiderjans, M. J. (1993). *Operations management*. Boston: Houghton Mifflin Company.

Markland, R. E., Vickery, S. K., & Davis, R. A. (1998). *Operations management: Concepts in manufacturing and services* (2nd ed.). St. Paul: South-Western College Publishing.

Norton, R. (1994). Which offices or stores really perform best? A new tool tells. *Fortune*, *130*(9), 38.

Troutt, M. D., Rai, A., Tadisina, S. K., & Zhang, A. (1998). A new efficiency methodology for IT investment analysis studies. In: M. A. Mahmood & E. J. Szewczak (Eds), *Measuring information technology investment payoff: Contemporary approaches* (pp. 202–222). Hershey, PA: Idea Group Publishing.

Van Zandt, G. M. (1998). *A production efficiency case study in a small organization utilizing efficiency ratio modeling*. Dissertation. Carbondale, IL: Southern Illinois University.

This page intentionally left blank

PART II:
MULTI-CRITERIA APPLICATIONS

This page intentionally left blank

A SURVEY OF MULTI-OBJECTIVE SCHEDULING TECHNIQUES APPLIED TO THE JOB SHOP PROBLEM (JSP)

David L. Bakuli

ABSTRACT

A survey of multi-objective scheduling techniques on the job shop problem is offered in this chapter. The survey traces the development of techniques from Integer programming to genetic algorithms that take advantage of the power of recent computing technology. Applications are in areas as diverse as job scheduling, nurse scheduling, and groundwater monitoring.

INTRODUCTION

A job shop is a production system for the assembly or manufacture of a variety of products where no single pattern or processing sequence or equipment is the same for all of the products. Various models incorporate different elements and different assumptions. For example, different assumptions can be made regarding the availability of jobs. It may be assumed that either there is a static pool of jobs to be processed or that jobs arrive dynamically according to a known probability distribution. Models

Applications of Management Science: In Productivity, Finance, and Operations
Applications of Management Science, Volume 12, 51–62

ISSN: 0276-8976/doi:10.1016/S0276-8976(06)12005-2

may incorporate the possibility of machine breakdowns, limitations on storage space, or buffers for work in progress or finished jobs. Each additional constraint further complicates the model.

However, for a job shop with a single machine, the problem is reduced to determining the order in which a set of jobs $\{J_1, J_2, \ldots, J_n\}$ will be processed by the machine in order to optimize some measure of performance. For example, the objective might be to minimize the total processing time of the jobs. The elegance of this problem arises from the fact that many processes can be modeled as job shops. For example, in centrally planned economies, and indeed in most local and state governments, a single department may be charged with the duty of managing a variety of development projects. Projects passing through such a department can be processed based on the job shop model. A developer faced with many construction projects has to sequence them subject to resource constraints. This system can also be modeled as a job shop. Thus, it is possible to model many problems with the job shop approach.

Most of the studies on the job shop problem (JSP) have aimed at optimizing a single objective. These measures of performance have included mean completion time, due date, number of tardy jobs, maximum tardiness, and inventory and utilization costs.

In reality, however, the optimization of a single objective implies that some other conflicting objective is left worse off than before. This inadequacy of traditional single objective optimization techniques to solve real time scheduling problems has led to a search for new approaches that incorporate the Decision Maker's (DM) preference function. Multiple objective optimization theory is based on the utility theory work of Von Neumann and Morgestein (1947), Keeney and Raiffa (1976), and many others. Although impressive in its approach, the theory has not been widely adopted by DMs according to the number of reported applications in industry.

This paper aims at reviewing some of the studies on the JSP that have used multiple objective approaches.

SINGLE OBJECTIVE OPTIMIZATION APPROACHES

From the traditional approaches to the job shop scheduling problem, we can infer that four types of information fully describe a specific job shop scheduling problem (Rochette, 1975):

(1) The jobs and their characteristics;
(2) The number and types of workers (and machines) that comprise the shop;

(3) The disciplines that restrict the manner in which assignments are made; and
(4) The criterion by which a schedule will be evaluated.

Most studies have aimed at optimizing a single objective such as, mean completion time, number of tardy jobs, maximum tardiness, etc. In real life, businesses deal with multiple objectives that conflict. For instance, a firm may have the following four objectives:

(1) maximize profit margin;
(2) minimize cost subject to a minimum level of equipment utilization;
(3) minimize power consumed by the plant; and
(4) minimize labor costs (without laying off any employees, in order to avoid industrial disputes).

Clearly, there is bound to be a conflict; it is neither beneficial to optimize a single objective at the exclusion of others, nor is it possible to have a single objective that when optimized will simultaneously optimize the other objectives.

One of the main conclusions of Panwalker, Dudek, and Smith (1973) in his research report on industrial scheduling was that managers do schedule according to multiple objectives. In spite of this finding, research has continued, unabated, to focus on single objective functions. What one finds is a series of algorithms, which do not optimize the entire system. The weighted tardiness performance measure has attracted most studies among the single objectives class. Works on this subject include Held and Karp (1962), Elmaghraby (1973), Lawler (1973), Emmons (1975), Baker (1975), and Martins (1984). The general approaches used in these studies are dynamic programming, implicit enumeration, and sometimes use of some elimination criteria. The number of jobs handled by these algorithms was limited to between 12 and 15. However, recent improvements in the storage and processing speed have improved this number to about 25 jobs (French, 1982). Moore (1968) developed an algorithm that minimizes the number of tardy jobs. Further development of this algorithm by Sidney (1973) allows that certain jobs have priority over others because they may not be delayed, hence in sequencing, they must be processed before their due dates.

The theme that runs through all the above scheduling problems is a desire to optimize a single objective. However, it should be noted that some of the performance measures are related. Optimizing a schedule with respect to one objective automatically optimizes any equivalent measures. For an enumeration of equivalent measures and their proofs refer to French (1982) and Conway, Maxwell, and Miller (1967).

Generally, the sequencing algorithms can be divided into two types, minimum makespan and due date:

(1) "Minimum makespan," if the objective is to complete all of the jobs in a static set as soon as possible; or
(2) "Due date," if the objective function aims at optimizing a value of some function of the time at which individual jobs are completed (Mellor, 1966).

None of these algorithms explicitly aims at optimizing more than one performance measure. The inability to find a common measure of value for the mixture of qualitative and quantitative goals has heightened interest in optimizing multiple objectives. Even though single objective scheduling problems remain intractable, there is increasing focus on multiple criteria optimizing techniques.

MULTI-OBJECTIVE OPTIMIZATION PROBLEMS

One of the criticisms leveled at the traditional approaches that use single optimization techniques has been that the analyst assumes that a DM does not have much to contribute to the final solution, apart from providing the required data on system operation. The analyst then classifies the problem into one of several predetermined categories for which a cookbook solution exists. While this may be a harsh criticism of the approaches, it is true, however, that the DM has had a low profile to play in the search for solutions to the business' problems largely in an effort to avoid his or her "subjective" opinions. This is the gap that multi-objective decision-making aims to bridge by integrating into the final solution the preference values of the DM. Although its use can be traced to as far back as the late nineteenth century when Vilfredo Pareto (1896) used it in welfare economics, it is not until the early half of the last twentieth century that the procedure was used in the operations management area. In particular, the works of von Neumman and Morgester (1947), and Kuhn and Tucker (1951) were instrumental in helping ground the multi-objective optimization theory. There is, however, not much correspondence between the economists' approach and that by the operations management researchers.

The general form of the multiple objective programming problem can be written as (Sawaragi, Nakagama, & Tanino, 1985):

$$\textit{Optimize}\, f(x) = (f_1(x), f_2(x), \ldots f_p(x)) \text{ over } x \in X \tag{1}$$

This formulation is also called a vector optimization problem. As we saw in a previous example, it may be required that some of the objectives that a firm may want to optimize must be maintained at certain levels, and if these are denoted by $g_j(x)$, then the constraints become:

$$g_j(x) \leq b_j \quad j = 1, 2, \ldots, m \tag{2}$$

Further, if there are technical constraints, $h_i(x)$, another set of constraints is added:

$$h_i(x) \leq b_i \quad i = 1, 2, \ldots, n \tag{3}$$

To arrive at a final solution, one does not aim at minimizing or maximizing a single representative objective function; rather aims at obtaining an acceptable balance in the achievement of each of the single objective. It is the DM who ultimately makes the selection of the preferred solution. Trade-offs have to be made, and this is where the importance of the utility theory comes into play providing the axioms needed to scale value judgments of the DM.

Various approaches are used in evaluation modeling. One approach aims at finding a scalar-valued function $u\,(f_1, f_2, \ldots, f_p)$ representing the DM's preference. Another approach is interactive modeling. In the latter approach, the analyst elicits information from the DM concerning his or her preferences. This information is either available at the beginning or can be obtained interactively by use of one of the decision analysis techniques. In the absence of such value preferences then the furthest the analyst could go is to generate a set of efficient solutions, which are then presented to the DM for selection of the one that he or she prefers. These are variously referred to as Non-dominated, Pareto optimal, or admissible solutions. Alternatively, at this point the traditional modeling techniques can be used to optimize the DM's preference function as:

$$\text{Max}\ u\,(f_1, f_2, \ldots, f_p)\ \text{over}\ x \in X \tag{4}$$

In the sense that the final solution is influenced by the value judgment of the DM, we cannot talk of an optimal solution as we do in the case of single objective optimization. The notion of efficiency is then introduced. A feasible solution is efficient if there is no other feasible solution with a higher value, i.e.

$$x \in X \text{ is efficient } \textit{iff} \text{ there is no } y \in Y \text{ such that } f(y) > f(x) \tag{5}$$

where f is the vector valued function whose ith component at x is $f_{\mathrm{i}}(x)$.

A number of algorithms have been developed to determine the whole set of the efficient solutions, some well-defined subset, or a close approximation to the efficient set. Hartley (1983) partitions the algorithms into three classes:

(1) A characterization of efficiency is used to transform the multiple objective problems into a family of single objective problems specified in terms of a number (usually p-1), of parameters. Well-known single optimization techniques coupled with parametric methods can then be used to find all solutions of this family.
(2) A standard single objective optimization method is modified to cope with multiple objective problems. It is necessary, of course, to establish conditions for the validity of such extensions. To date the most favored technique for this treatment is dynamic programming.
(3) The final class contains all methods, which do not correspond to any standard single objective method. Clearly, any such method can be specialized to the single objective case and, therefore, the presumption would be that the specialization is inferior, in this case to standard methods. However, if the reason for this is that computation time is not as heavily dependent on the number of objectives as in other methods, a compensating advantage may be that the approach becomes relatively more attractive with a large number of objective functions.

EFFICIENT SCHEDULES

Van Wassenhove and Gelders (1980) were the first to introduce the concept of efficiency into the context of scheduling. They considered a single machine with two objectives: the vector minimization of mean flow time and maximum tardiness.

The algorithm developed is a modification of the earlier work by Smith (1956), which solved the problem of minimizing F subject to the constraint that $T_{\max} = 0$. This problem was modified by Heck and Roberts (1972) on the same fashion as Van Wassenhove and Gelders (1980), but they did not introduce the notion of efficiency. In this algorithm, ideas of goal programming are introduced into scheduling. The problem structure is such that the data requested is in integers. It is possible, as French (1983) observes, to modify it so that non-integer data is acceptable. Their algorithm, however, only finds a representative set of the efficient schedules and even then, it has implicit enumeration at step 1, which has high-computational implications. The authors point out that it is possible to modify the algorithm to find all the efficient schedules if so desired.

Another approach in this group of finding efficient scheduling is by Van Wassenhove and Baker (1980). The philosophy used in the second approach is that processing times are not fixed but variable and in reality it is possible to expedite jobs at an extra cost, either by use of extra resources, overtime or a combination of them. Indeed we find this same assumption in Optimized Production Technology (OPT) where we have substitute resources when needed. Using this approach, two greedy algorithms are developed.

The processing times are set to lie in the interval $a_i < t_i < b_i$ for $i = 1, 2, \ldots, n$. In this range they can select any t_i at a cost of $c_i(b_i - t_i)$. The total cost of processing is

$$\sum_{i=1}^{n} c_i(b_i - t_i) \tag{6}$$

When $t_i < b_i$, job j_i is said to have been crashed by $(b_i - t_i)$.

Van Wassenhove and Baker's algorithm solves min $(T_{\max}, S)$ for a single machine problem. The general assumption made is that the jobs are already sequenced in order of non-decreasing due date. What the algorithm does is to trace out an efficient frontier in $(T_{\max}, S)$ space. Their second algorithm solves a more general case but first sequences the jobs using Lawler's (1973) algorithm before finding an efficient set. The general greedy algorithm's computational complexity is of order $O(n^3)$ and as such is not computationally efficient.

INTEGER PROGRAMMING APPROACHES

The structure of many scheduling problems makes it possible for them to be formulated as multiple objective mixed integer linear programming problems. For a survey of these formulations refer to Rinnooy (1976). Huckert et al. (1980), have used a multiple objective integer programming formulation due to Manne (1960). In their formulation they have used an interactive approach that enables the DM to choose a subset of the efficient points set in the frontier that is of interest to him. The computational requirements of the problem are reduced.

After the problem has been recast as an integer program (IP), known single objective optimization techniques are used to find a solution. We should note here that IP techniques are not tractable with single objective problems, so there is not much to be gained from this approach. In fact Huckert et al formulated a 20-job, 10-machine problem with 1,900 binary variables, 201 continuous variables, and 4,000 constraints, which could not be solved by present day computers.

MULTIPLE-ATTRIBUTE VALUE FUNCTION APPROACHES

The other approach to multiple scheduling problems is to find an optimal schedule with respect to a weighted Multi-attribute value function. Allowance is made for trade-offs between conflicting objectives. Vickson (1980) investigated a single machine problem in which the processing times are variable. See the formulation by Van Wassenhove and Baker (1980). The objective function is a weighted sum of processing costs and weighted mean flow time:

$$TC = w_1 \sum_{i=1}^{n} C_i(b_i - t_i) + w_2 \sum_{i=1}^{n} \alpha_i F_i \tag{7}$$

Vickson did not find a polynomial time algorithm for this problem. He observes that, "It does not seem to be theoretically easy in the present generalized form, but its actual computational complexity status is not yet determined." However, he did come up with a branch and bound technique and a heuristic of outstanding performance. The heuristic only failed to produce the optimal solution in one problem for which it gave a solution only 0.01% above the optimum. Perhaps these results are not surprising since the structure of the problem is such that if we already know the sequence, then the objective function is merely a weighted average of the processing times. Other complications such as getting an initial feasible schedule are also not present.

Norbis (1987) studied the Resource Constrained Scheduling problem using a 0–1 Multiple objective approach. He reduced seven objective functions into three broad categories:

(1) Completion time related;
(2) Due date related; and
(3) Resource utilization related.

The optimization was subject to constraints on resource availability. He came up with a heuristic that performs quite well.

In this category, however, it is Kao (1980) who has the most general approach. He considers an n-job, single machine problem in which the objective is to minimize an additive value function:

$$u(x_1,\ x_2,\ldots,x_m)=\sum_{i=1}^{m}u_m(x_m) \tag{8}$$

With $x_1,\ x_2,\ldots,x_m$ being measures of the performance e.g. $T_{\max}$, m, F, etc. and u_m(.)are monotonic decreasing single attribute Value Functions with the assumption that they are mutually utility independent. It is possible to extend this method to a non-additive value function under certain conditions.

Kao uses a combination of branch and bound and dynamic programming techniques to generate an efficient schedule. At each node, m single objective problems are solved. This makes the algorithm time-consuming. Again the use of dynamic programming places high demand on the storage capacity of the computer although, it implicitly eliminates some nodes.

RECENT APPROACHES

More recent approaches to the JSP have considered the use of genetic algorithms; simulated annealing and tabu search. Genetic algorithms use mechanisms similar to those believed to apply in natural genetics. The algorithms are based on the idea of generating new solutions from an initial set of parent solutions. Table 1 summarizes the paper using these approaches. Cochran, Horng, and Fowler (2003) consider three objectives:

(1) Minimizing makespan;
(2) Minimizing total weighted completion time (TWC);
(3) Minimizing total weighted tardiness (TWT).

The authors observe that, "simultaneous control of the above performance measures should lead to a superior schedule. Such as a schedule will increase throughput, reduce inventory holding costs, reduce cycle time, and lead to on-time delivery."

Setamaa-Karkainen, Miettinen, and Vuori, (2005) introduce a network connection selection problem and formulate it as a multi-objective scheduling problem where the objective is to minimize both costs and connection time.

Table 1. Recent Multi-Objective Approaches.

Method	Objectives	Author(s)
Goal programming (GP)	Pareto efficient set	Lee and Jung (1989)
Genetic algorithm (GA)		Hyun, Kim, and Kim (1998) Chen, Narita, Tsuji, and Sa (1996)
	-	Omori, Sakakibara, and Suzuki (1997) Jaskiewicz (1997) Cieniawski, Eheart, and Ranjithan (1995)
Vector evaluated genetic algorithm (VEGA)	Makespan, TWC, TWT	Schaffer (1985)
Multi-objective genetic algorithm (MOGA)		Murata, Ishibuchi, and Tanaka (1996)
Two-stage multi-population genetic algorithm (MPGA)		Cochran et al (2003)
Simulated annealing and tabu search		Marret and Wright (1996), Jaskiewicz (1997)
Petri nets		Yim and Lee(1996)

REFERENCES

Baker, K. R. (1975). A comparative survey of flow-shop algorithms. *Operations Research*, *23*, 62–73.

Chen, Y., Narita, M., Tsuji, M., & Sa, S. (1996). A genetic algorithm approach to optimization for the radiological worker allocation problem. *Health Physics*, *70*, 180–186.

Cieniawski, S. E., Eheart, J. W., & Ranjithan, S. (1995). Using genetic algorithms to solve a multiobjective groundwater monitoring problem. *Water Resources Research*, *31*, 399–409.

Cochran, J. K., Horng, S.-M., & Fowler, J., W. (2003). A multi-population genetic algorithm to solve multi-objective scheduling problems for parallel machines. *Computers & Operations Research*, *30*, 1087–1102.

Conway, R. W., Maxwell, W. L., & Miller, L. W. (1967). *Stheory of scheduling*. Reading, MA: Addison-Wesley.

Elmaghraby, S. E. (1973). *Symposium on the theory of scheduling and its applications.* New York: Springer.

Emmons, H. (1975). A note on a scheduling problem with dual criteria. *Naval Research Logistics Quarterly*, *22*(3), 615–616.

French, S. (1982). *Sequencing and scheduling: An introduction to the mathematics of the job-shop*. England: Ellis Horwood Ltd.

Hartley, R. (1983). Survey of algorithms for vector optimization problems. In: S. French, L. C. Thomas, R. Hartley & D. J. White (Eds), *Multi-objective decision making*. London; New York: Academic Press.

Heck, R., & Roberts, S. (1972). A note on the extension of a result on scheduling with secondary criteria. *Naval Research Logistics Quarterly*, *19*, 403–405.

Held, M., & Karp, R. M. (1962). A dynamic programming approach to sequencing problems. *Journal SIAM*, *10*, 196–210.

Huckert, K., Rhode, R., Roglin, O., & Weber, R. (1980). On the interactive solution of a multicriteria scheduling problem. *Operations Research*, *24*, 47–60.

Hyun, C. J., Kim, Y., & Kim, Y. K. (1998). A genetic algorithm for multiple objective sequencing problems in mixed model assembly. *Computers & Operations Research*, *25*, 675–690.

Jaskiewicz, A. (1997). A metaheuristic approach to multiple objective nurse scheduling. *Foundations of Computing and Decision Sciences*, *22*, 169–183.

Kao, E. P. C. (1980). A multiple objective theoretic approach to one-machine scheduling problems. *Computers and Operations Research*, *7*, 251–259.

Keeney, R. L., & Raiffa, H. (1976). *Decision with multiple objectives: Preferences and value tradeoffs*. NY: Wiley.

Kuhn, H. W., & Tucker, A. W. (1951). Nonlinear programming. In: *Proceedings of the 2nd Berkeley symposium on mathematical statistics and probability*, Berkeley, CA (pp. 481–492).

Lawler, E. L. (1973). Optimal sequencing of a single machine subject to precedence constraints. *Management Science*, *19*, 544–546.

Lee, S. M., & Jung, H.-J. (1989). A multi-objective production planning model in a flexible manufacturing environment. *International Journal of Production Research*, *27*, 1981–1992.

Manne, A. S. (1960). On the job – shop scheduling problem. *Operations Research*, *8*, 219–223.

Marret, R., & Wright, M. (1996). A comparison of neighborhood search techniques for multi-objective combinatorial problems. *Computers and Operations Research*, *23*, 465–483.

Martins, E. Q. V. (1984). On a special class of bicriterion path problems. *European Journal of Operational Research*, *17*, 85–94.

Mellor, P. (1966). A review of job-shop scheduling. *Operational Research Quarterly*, *17*, 161–171.

Moore, J. M. (1968). An n-job, one machine sequencing algorithm for minimizing the number of late jobs. *Management Science*, *15*, 102–109.

Murata, T., Ishibuchi, H., & Tanaka, H. (1996). Multi-objective genetic algorithm and its application to flowshop scheduling. *Computers and Industrial Engineering*, *30*, 957–968.

Norbis, M. (1987). *Heuristics for the resource constrained scheduling problem.* Unpbublished Doctoral Dissertation, University of Massachusetts, Amherst, MA.

Omori, R., Sakakibara, Y., & Suzuki, A. (1997). Applications of genetic algorithms to optimization problems in the solvent extraction process for spent nuclear fuel. *Nuclear Technology*, *118*, 26–31.

Panwalker, S. S., Dudek, R. A., & Smith, M. L. (1973). Sequencing research and the scheduling problem. In: S. E. Elmaghraby (Ed.), *Symposium on the theory of scheduling and its applications*. New York: Springer-Verlag.

Pareto, V. (1896). *Cours d'Economique, Vols. 1 and 2, F.* Rouge: Lausanne, Switzerland.

Rinnooy Kan, A. H. G. (1976). *Machine scheduling problems*. Martinus Nijhoff: The Hague.

Rochette. (1975). *A statistical analysis of the job shop scheduling problem.* Unpublished Ph.D. dissertation, University of Massachusetts, Amherst, MA, USA.

Sawaragi, Y., Nakagama, H., & Tanino, T. (1985). *Theory of multi-objective optimization.* New York, NY: Academic Press, Inc.

Schaffer, J. D. (1985). Multiobjective optimization with vector evaluated genetic algorithms. *Proceedings of the 1st ICGA*, (pp. 93–100).

Setamaa-Karkkainen, A., Miettinen, K., & Vuori, J. (2005). *Best compromise solution for a new multiobjective scheduling problem*, Computers & Operations Research, Available online at http://www.sciencedirect.com

Sidney, J. B. (1973). An extension of Moore's due date algorithm. In: S. E. Elmaghraby (Ed.), *Symposium on the theory of scheduling and its applications*. Berlin, New York: Springer-Verlag.

Smith, W. E. (1956). Various optimizers for single stage production. *Naval Research Logistics Quarterly*, *3*, 59–66.

Van Wassenhove, L. C. & Baker, K. R. (1980). A bicriterion approach to time/cost trade-offs in sequencing. 4th European Congress on Operational Research, Cambridge, England, July 22–25.

Van Wassenhove, L. C., & Gelders, L. F. (1980). Solving a bicriterion scheduling problem. *European Journal of Operational Research*, *4*, 42–48.

Vickson, R. G. (1980). Choosing the job sequence and processing times to minimize total processing time plus flow cost on a single machine. *Operations Research*, *28*, 1155–1167.

Von Neumann, J., & Morgestern, O. (1947). *Theory of games and economic behavior* (2nd edn.). Princeton, NJ: Princeton University Press.

Yim, S. J., & Lee, D. Y. (1996). Multiple objective scheduling for flexible manufacturing systems using Petri nets and heuristic search. *Proceedings IEEE International Conference on Systems, Man and Cybernetics*, *4*, 2984–2989.

A MULTI-OBJECTIVE MATHEMATICAL PROGRAMMING MODEL FOR AUDIT SAMPLES OF BALANCES FOR ACCOUNTS RECEIVABLE

Kenneth D. Lawrence, Ronald Klimberg and Sheila M. Lawrence

ABSTRACT

This paper will detail the development of a multi-objective mathematical programming model for audit sampling of balances for accounts receivable. The nonlinear nature of the model structure will require the use of a nonlinear solution algorithm, such as the GRG or the genetic algorithm embedded in a Solver spreadsheet modeling system, to obtain appropriate results.

INTRODUCTION

Auditing

Large firms with rapidly expanding masses of financial data are forced to test-check in the audit before they really had a chance to fully assess all

Applications of Management Science: In Productivity, Finance, and Operations
Applications of Management Science, Volume 12, 63–76

ISSN: 0276-8976/doi:10.1016/S0276-8976(06)12006-4

implications. Whenever an auditor decides that testing is an appropriate audit procedure, statistical sampling is always appropriate. If non-statistical methods were chosen, then the auditor would need to choose between an under-auditing or over-auditing procedure (Fienberg, Neter, & Leitch, 1977; Hall, Hunton, & Pierce, 2000, 2002).

Owing to limited audit resources, it is virtually impossible to examine all transactions as a means of determining that the stockholders' interests are being adequately protected. It is obvious that some testing or sampling is the only practical method for drawing conclusions about the condition of all financial transactions. The auditor needs only an acceptable degree of assurance that the findings are accurate.

When statistical sampling is used, the auditor can calculate the risk associated with the use of sample data in lieu of a complete audit of all financial records. Random selection allows for the measurement of risk through probability. Without randomness, there is no way to measure the sampling risk. In using statistical sampling, the auditor can select financial transactions in an objective manner. The auditor's professional judgment is neither replaced nor diminished by the use of statistical sampling. Statistical sampling is a refinement of the auditor's test/checking procedure.

Statistical sampling provides the auditor with a procedure for calculating risk and precision associated with the test results. All audit efforts are directed toward the fulfillment of specific audit objectives. The primary objective of auditing is the issuance of a financial report, which reflects the auditor's opinion.

The advantages of statistical sampling in an audit are as follows:

- The amount of test does not increase in proportion to the increased size of the universe tested.
- The sample selection is more objective and defined.
- It provides a device for calculated risk.
- It provides an accurate decryption of large sets of data.
- It provides better audit coverage.
- It provides consistency.
- It provides better reliability of results.
- It is an efficient and effective procedure.

The Audit Sample of Accounts Receivable

The objective of auditing the sales and collection process is to evaluate whether sales, returns, allowance, bad debt, etc., are properly reflected in the firm's financial records in accordance with appropriate accounting standards. The tests of transactions enable the auditor to decide on the appropriately assessed level of control risk and substantive tests of transactions to evaluate the correctness of the dollar accounts of the financial records (American Institute of Certified Public Accountants, 1997; Gary & Johnson, 1979; Felix & Grimlund, 1977; Horgan, 1997; Kraft, 1968; Neter, Leitch, & Fienberg, 1978; Teitlebaum & Robinson, 1975).

Tests of control procedures for determining a firm's internal control effectives need to be performed when the auditor intends to reduce assessed control risk to the planned level and thereby reduce planned substantive tests.

Three major factors affect the number of sample items for tests of transactions:

- The deviation rate the auditor is willing to accept in the population.
- The deviation rate expected in the sample.
- The risk the auditor is willing to take that the sample will not be representative of the population.

The standard objectives of an audit to obtain an accurate estimate of account balances related to audit objectives include:

- The accounts receivable in the aged trial balance agrees with the related master file accounts and agrees with the general ledger.
- Recorded accounts receivables exist.
- Existing accounts receivable are included.
- Accounts receivable are accurate.
- Accounts receivable are properly classified.
- Cut off for accounts receivable is correct.
- Accounts Receivable is stated at realizable values.

The major factors affecting sample size for confirming accounts receivable fall into the following categories:

- tolerable misstatement;
- inherent risk (relative size of total accounts receivable), (number of accounts), (prior year results), (expected misstatements);
- control risk;
- achieved detection risk from other substantive tests;
- type of confirmation.

AUDIT SAMPLING

Representative Sampling

A representative sample is one in which the characteristics in the sample audit approximate the population characteristics. This means that those items sampled are similar to the items not sampled in the audit. In practice, it is difficult to know how representative the sample is. However, an auditor can increase the likelihood of a sample's being representative by using extreme care in the design, selection, and evaluation of the sample.

The auditor is concerned about two types of risks from the sampling: non-sampling and sampling risk. The sampling risk is due to some bias in the sample collection resulting in a non-representative of the population. In such a case, this situation could lead to the auditor's making incorrect conclusions. Sampling risk can be controlled by using an appropriate method of selecting sample items from the population and by adjusting the sample size. If at all possible, sampling risk should be eliminated. The second risk is the non-sampling risk' and it is the risk that the audit does not uncover existing exceptions in the sample. This risk is the expected statistical risk from taking a representative sample (as opposed to sampling the entire population). Careful design of audit procedures, proper instruction, supervision, and review are ways to control the non-sampling risk.

In order to obtain an appropriate representative sample, we need to estimate a population parameter with a given confidence level and degree of estimation precision. In the typical auditing process, the auditor estimates an error rate within a population of the dollar value of the items.

In particular, for a stratified sampling example, with *I* stratum and *J* territories within each stratum, let us define and assume that we have estimates of the following parameters:

$p_i =$ estimated probability of error in stratum i
$w_i =$ the proportion of items in the ith stratum.

Additionally, let us define the unknown audit sample size and variance as:

$x_{ij} =$ the number of samples collected in ith stratum and territory j; $(i = 1, 2, \ldots, I)$;
$(j = 1, 2, \ldots, J)$;
$v_i =$ the variance of the estimate of the error rate of the i stratum.

Mathematically, the estimate of the variance of the *i*th stratum is:

$$v_i = \frac{[w_i^2 p_i(1 - p_i)]}{\sum_{j=1}^{J} x_{ij}}$$

The auditor seeks to find a balance between the audit sample size and the size of the variance.

Stratification in audit sampling is most desirable. A typical approach to stratification is to consider both the dollar size of the individual accounts and the length of time an account has been outstanding as a basis for selecting the balances for information. It is quite common in most audits that the confirmation of older and larger balances is of primary consideration since they are the most likely to have the most significant misstatement. However, it is also important to sample items from every material segment of the audit population. Additionally, many auditors select all accounts above a certain dollar amount and select random samples from the remainder (Arkin, 1961; Baker & Copeland, 1979; Bonner, 1999; Elliott & Rogers, 1972; Hitzig, 1998; Loebbeck & Neter, 1975).

Dollar Units Sampling

Dollar unit sampling can be reviewed as an extension of proportionate sampling. When the sampling universe is stratified by dollar value, a method of allocation is to sample in proportion to the dollar value of the transaction in each stratum. If this approach is carried to the point where each transaction forms a stratum, then the result would be dollar unit sampling. This dollar unit sampling optimizes the efficiency obtained through a dollar stratification. Thus, it eliminates problems associated with determining strata boundaries and in allocating the sample among the strata (Anderson & Teitlebaum, 1973; Garstka, 1977; McCray, 1973; Teitlebaum & Schwartz, 1958).

Generally dollar unit sampling will involve only one or two explicit strata. The number of strata may be kept small by reporting the accounts into broad categories such as above average, average, and below average.

In order to perform dollar unit sampling, an estimate of the error variance in each stratum is needed. For each cell (i, j) there may be many items that have no errors. The other items will have errors of varying amount about the reported dollar value. An assumption will be that the error distribution has a mean of zero and it is symmetrical. Moreover, the standard deviation of the error distribution is proportional to the average dollar value D_i of the cell. Thus, we can define the dollar unit error variance of each cell as: $u_{ij} = w_i^2 p_i (fD_i)^2 / x_{ij}$, where f is a constant of proportionality.

Attribute Sampling

Attribute sampling is based on the binomial distribution. The binomial distribution is a probability distribution of all possible independent samples when each of the items in the population has two possible states (i.e., yes/no, defect/non-defect, and errors/no errors). In practical audit situations, auditors do not take repeated samples from known populations. They take a sample from an unknown population and get a specific number of exceptions in that sample. Attribute sampling allows for the desire to correct as many errors as possible during the audit process. Attribute sampling involves the estimate of an overall error rate. The attribute sampling process requires the auditor to specify an expected occurrence rate in order to determine the needed sample size. If there is an over-estimate, then the audit sample will be over-sampled. In order to determine the minimum sample size, the auditor needs to specify the maximum tolerable occurrence rate. Typically sampling for attributes is considered prior to sampling for variables.

Given the population size of cell (i, j), n_{ij}, the expected number of errors remaining is defined as: $\sum_{i=1}^{I}\sum_{j=1}^{J}p_i(n_{ij} - x_{ij})$.

A MULTI-OBJECTIVE NONLINEAR MODEL FOR AUDIT SAMPLING

Our model for audit sampling of balances for accounts receivable is a nonlinear sequential multiple objective programming model (Arthur & Lawrence, 1985; Ashton & Atkins, 1979; Bhaskar & McNamee, 1979; Booth & Dash, 1977; Chateau, 1975; Coffin & Taylor, 1996; Fowler & Schniederjans, 1987; Giokas & Vassiloglou, 1991; Goedhart & Spronk, 1995; Gonzalez, Reeves, & Franz, 1987; Guerard & Stone, 1987; Keown & Martin, 1976; Lawrence & Marose, 1993; Sheshai, Harwood, & Hermanson, 1977; Rashid & Tabucanon, 1991). The mathematical formulation of this problem is as follows:

$$\text{Min } Z_1 = \sum_{i=1}^{I}\sum_{j=1}^{J} x_{ij} \qquad (1)$$

$$\text{Min } Z_2 = \text{MIN}\left[\operatorname*{Max}_{i}(v_i)\right] = \text{MIN}\left[\operatorname*{Max}_{i}\left(\frac{[w_i^2 p_i(1-p_i)]}{\sum_{j=1}^{J} x_{ij}}\right)\right] \qquad (2)$$

$$\text{Min } Z_3 = \text{MIN}\left[\underset{ij}{\text{Max}}(u_{ij})\right] = \text{MIN}\left[\underset{ij}{\text{Max}}\left(\frac{w_i^2 p_i (fD_i)^2}{x_{ij}}\right)\right] \tag{3}$$

$$\text{Min } Z_4 = \sum_{i=1}^{I}\sum_{j=1}^{J} p_i(n_{ij} - x_{ij}) \tag{4}$$

s.t

$$L_{ij} \leq x_{ij} \leq U_{ij} \quad \forall i, j \tag{5}$$

$$\sum_{i=1}^{I}\sum_{j=1}^{J} x_{ij} \leq M \tag{6}$$

$$x_{ij} \geq 0 \quad \forall i, j$$

The model takes into consideration the intent of the three different sampling plans – representative, dollar units, and attribute sampling. The first objective, (1), minimize the total audit sample size. The second objective, (2), minimizes the maximum variance of the error rate of the stratums. The third objective, (3), minimizes the maximum dollar unit error variance of each cell *ij* (each stratum/territory). The last objective, (4), minimizes the number of errors remaining. The first constraint, (5), provides upper and lower bounds, U_{ij} and L_{ij}, respectively, on the sample size for each stratum/territory. The last constraint, (6), requires the total audit sample size to be less than an upper bound, *M*.

The model structure will require the use of a genetic type algorithm in conjunction with a Premium Solver found in an Excel Solver. Given that the model is a multiple objective sequential nonlinear model, the solution of it would be greatly helped by the Premium Solver and its evolutionary (genetic) algorithm. A genetic algorithm starts with a set of chromosomes (numerical vectors), which represent the possible solutions to the optimization problem. The individual components within a chromosome are referred to as genes. New chromosomes are created by crossover and mutations. Crossover is the probabilistic exchange of values in a solution vector. Mutation is the random replacement of values in the solution vector. Chromosomes are then evaluated according to the objective function (fitness for surviving into the next generation). The result is a gene pool that evolves over time to produce better and better solutions to the given problem.

To a certain extent, Solver's evolutionary algorithm picks up where the nonlinear GRG algorithm leaves off. The solution that the Solver generates depends on the starting point and may be a local rather than a global optimum.

Solver tends to have difficulty solving problems with discontinuities and unsmooth landscapes. Although the evolutionary solver cannot completely avoid the possibility of becoming rapped at local optimum solution, its use of a randomized initial guess pool probabilistic crossover, and mutation operators make this less likely (Collins & Golden, 1988; Gary & Johnson, 1979; Glover & Greenberg, 1989; Goldberg, 1989; Holland, 1975; Johnson, Aragon, McGeoch, & Schevon, 1989; Lundy & Mees, 1986).

In addition, evolutionary algorithms can operate on any operational model.

AN EXAMPLE PROBLEM

An example with four strata, each containing four territories, was created and used to test the nonlinear multi-objective mathematical programming model for audit sampling. Tables 1 and 2 list the necessary input parameter values. Associated with each stratum are the following: an error rate, a proportion of items, and an average dollar value, listed in Table 1. The additional input parameter values, listed in Table 2, are the lower and upper bounds on the sample size of each stratum/territory.

The model was solved sequentially for each objective using the nonlinear GRG algorithm. Three out of the four objective solutions did not change after the first objective solution, e.g., once the third objective (minimizes the maximum dollar unit error variance) was solved, the following solutions, solved sequentially and including the other objectives, did not change from the initial solution (Table 3 lists and defines the four objectives). Nevertheless, these initial objective function solutions did vary from one objective function to another as shown graphically in Fig. 1. In particular in Fig. 1, the first left-most column shows the objective function values when only the first objective (one) was optimized; the second left-most column shows the

Table 1. Input Parameter Values.

Stratum	Error Rate(*Pi*)	Proportion(*wi*)	Average Dollar Value (*Di*)
1	0.015	0.35	75
2	0.025	0.30	30
3	0.020	0.15	94
4	0.035	0.10	82

Table 2. Input Parameter Values.

Stratum	Territory	Lower Bound	Upper Bound
1	1	50	500
1	2	90	900
1	3	60	600
1	4	50	500
2	1	100	1,000
2	2	80	800
2	3	70	700
2	4	50	500
3	1	60	600
3	2	50	500
3	3	25	250
3	4	15	150
4	1	40	400
4	2	30	300
4	3	20	200
4	4	10	100

Table 3. List of the Model's Objectives and their Definitions.

Objective	
1	Minimize sample size
2	Minimize the maximum variance of the error rates of the stratum
3	Minimize the maximum dollar unit error variance of each stratum/territory
4	Minimize the number of errors remaining

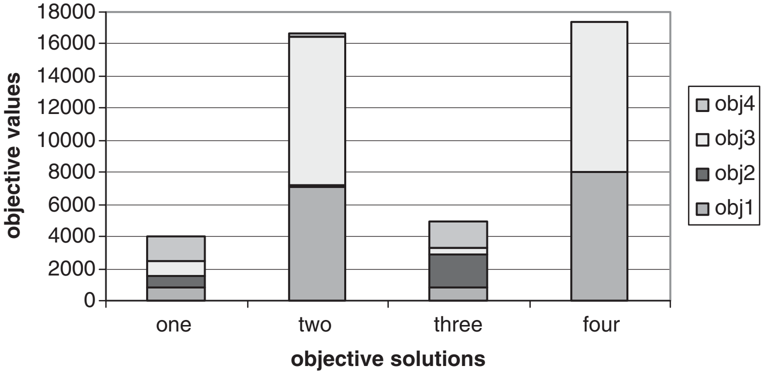


Fig. 1. Initial Objective Values for the Four Objective Solutions Using NLP.

objective function values when only the second objective (two) was optimized; and so on.

An initial obvious result, shown in Fig. 1, is that when a particular objective function was initially optimized, it achieved its minimum value. Next, the value of objective one is the overall sample size. This objective is at its lowest possible value of 800 in objective solutions one and three. In the other two objective solutions, two and four, the overall sample sizes escalate to 7,100 and 8,000, respectively (for this example, 8,000 is the largest possible overall sample size). Additionally, as the sample size increases in these two objective solutions, two and three, the value of objective three increases significantly while the value of objective four decreases.

The one objective in which the solutions did change when sequentially solved, using the nonlinear GRG algorithm, was the second objective to minimize the maximum variance of the strata error rates. Fig. 2 illustrates these solution changes. The rows in Fig. 2 show the different solutions as each objective is added sequentially to the model, i.e., the first row simply minimizes objective two; the second row minimizes objective three, given the

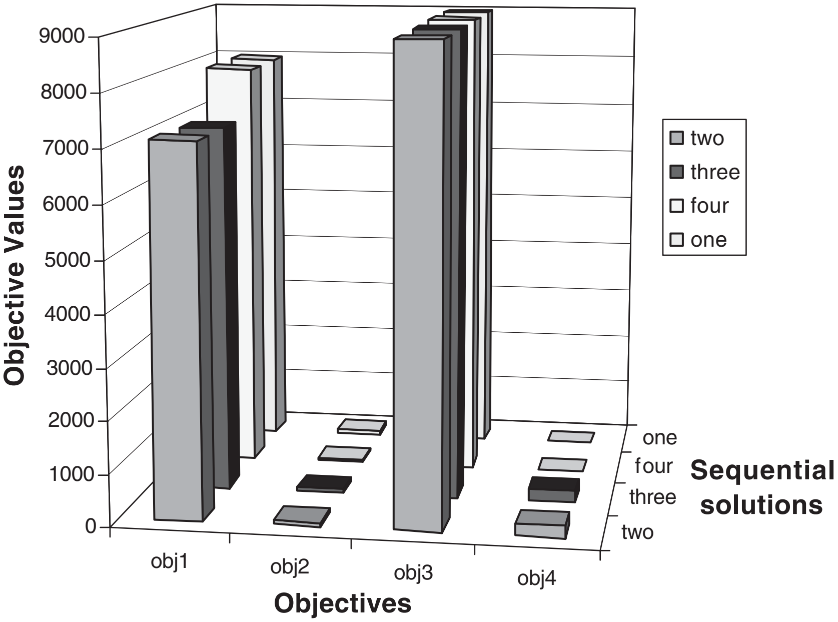


Fig. 2. Objective Two Sequential Solution Changes.

value of objective two in the first row, and so on. A particular row of bars in Fig. 2 illustrates the values of each of the objective functions, i.e., the first row in Fig. 2, where just objective two is minimized, is the same as the second left-most column in Fig. 1.

The nonlinear multi-objective mathematical programming for audit sampling was also solved sequentially using Solver's evolutionary algorithm. The evolutionary algorithm gave identical solutions as the NLP algorithm for the first, third, and fourth objectives. As with the NLP algorithm, the sequential solutions to the second objective varied using the evolutionary algorithm. However, these sequential solutions to the second objective, using the evolutionary algorithm, were different from the sequential solutions to the second objective using the NLP algorithm as shown in Table 4.

Table 4 lists the objective function values and their differences for the sequentially solved models using the NLP and evolutionary algorithms. (Objective two was first solved for; then objective three, four, and one.) Each of these eight solutions varied. No algorithm consistently out-performed any other, i.e., sometimes the NLP performed better, and sometimes evolutionary algorithm performed better. In fact, the NLP algorithm has five objective function values better than its corresponding evolutionary algorithm objective function value. On the other hand, the evolutionary algorithm has seven better objective function values.

Lastly, we applied the multi-objective weighting method and solved simultaneously the nonlinear multi-objective mathematical programming model for audit sampling. We solved the model for various weights, taking two approaches: (1) apply equal weights to the four objectives (we called this solution sum), and (2) assign a weight to the first objective and then equally disperse the remaining weights among the other objectives (we identify these solutions by the weight assigned to the first objective). Fig. 3 shows the

Table 4. Objective Function Values Sequentially Solved for the Second Objective using NLP and Evolutionary Algorithms.

	Objective Solved											
	Two			Three			Four			One		
	NLP	Evol.	Diff	NLP	Evol.	Diff	NLP	Evol.	Diff	NLP	Evol.	Diff
Obj1	7100.0	6956.5	143.5	7075.1	6974.9	100.2	7975.1	7955.6	19.6	7975.1	7955.6	19.6
Obj2	73.1	73.1	0.0	73.1	73.1	0.0	73.1	73.1	0.0	73.1	73.1	0.0
Obj3	9302.3	9284.1	18.3	9045.3	9215.4	−170.1	9045.3	9215.3	−170.1	9045.3	9215.3	−170.1
Obj4	225.0	261.8	−36.8	228.7	249.0	−20.2	3.7	13.9	−10.1	3.7	13.9	−10.1

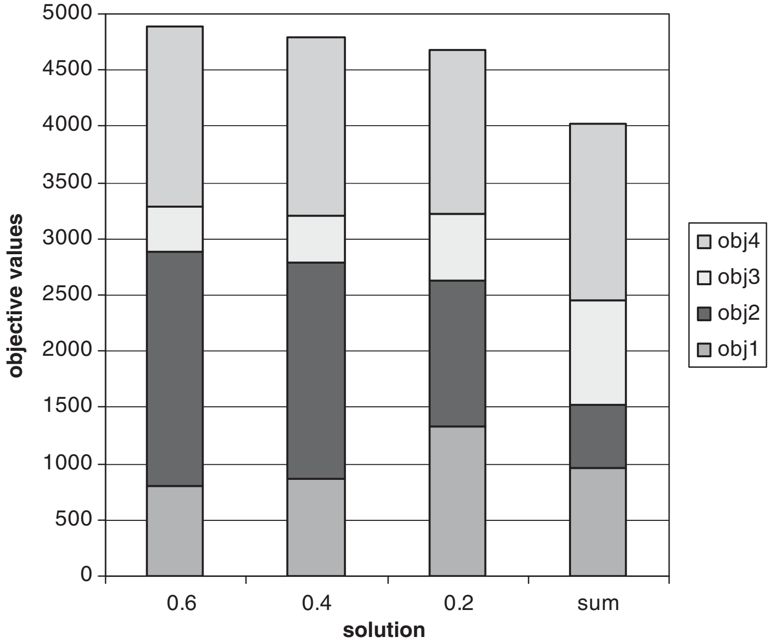


Fig. 3. Objective Function Values when Using the Weighting Method.

model results for equal weights (sum) and various weights assigned to the first objective. As illustrated in Fig. 3, a variety of tradeoffs occur as the weights are varied. Finally, the solutions shown in Fig. 3 were from using the NLP algorithm. In this case, when the evolutionary algorithm was applied, only occasionally relatively small differences occurred.

CONCLUSIONS

This paper has presented a multi-objective approach to audit sampling for balances of accounts receivable. Such a modeling approach will allow the auditor in a firm to develop an audit sample in which the financial records of the firm produce a highly representative view of the firm's sales, returns, allowances, bad debts, etc. This approach minimizes the total audit sample size, the maximum variance of the error variance, the maximum dollar unit error variance, and the number of errors. This is a far superior approach to developing an audit sample, as compared with traditional audit sampling techniques.

REFERENCES

American Institute of Certified Public Accountants. (1997). *Codification of statements on auditing standards*. New York: AICPA.

Anderson, R., & Teitlebaum, A. (1973). Dollar unit sampling: A solution to the audit sampling dilemma. *CA Magazine*, (April), 30–39.

Arkin, H. (1961). Discovery sampling in auditing. *Journal of Accountancy*, *12*(2), 101–107.

Arthur, J. L., & Lawrence, K. D. (1985). A multiple goal capital flow model for a chemical and pharmaceutical company. *The Engineering Economist*, *30*(2), 121–134.

Ashton, D. J., & Atkins, D. R. (1979). Multicriteria programming for financial planning. *Journal of the Operational Research Society*, *30*(3), 259–270.

Baker, R., & Copeland. (1979). Evaluation of the stratified regression estimator for auditing accounting populations. *Journal of Accounting Research*, *17*(2), 606–617.

Bhaskar, K., & McNamee, P. (1979). A multiple objective approach to capital budgeting. *Accounting and Business Research*, *10*(37), 25–46.

Bonner, S. (1999). Judgment and decision making research in accounting. *Accounting Horizons*, *13*(4), 385–398.

Booth, G. G., & Dash, G. H. (1977). Bank portfolio management using non-linear goal programming. *The Financial Review*, *14*(1), 59–69.

Chateau, J. P. D. (1975). The capital budgeting problem under conflicting financial policies. *Journal of Business Finance and Accounting*, *2*(2), 38–103.

Coffin, M. A., & Taylor, B. W. (1996). Multiple criteria R&D project selection and scheduling using fuzzy logic. *Computers and Operations Research*, *23*(3), 207–220.

Collins, E., & Golden, B. (1988). Simulated annealing bibliography. *American Journal of Mathematical and Management Sciences*, *8*, 211–307.

Elliott, R., & Rogers, J. (1972). Relating statistical sampling to audit objectives. *Journal of Accountancy*, *23*(7), 691–696.

Felix,, W. 1,, & Grimlund, R. (1977). A Sampling Model for Audit Tests of Composite Accounts. *Journal of Accounting Research*, *38*(2), 282–290.

Fienberg, S., Neter, J., & Leitch, R. (1977). Estimating the total overstatement error in accounting populations. *Journal of the American Statistical Association*, *72*(358), 295–302.

Fowler, K. L., & Schniederjans, M. J. (1987). A goal programming model for strategic acquisition problem solving. In: K. D. Lawrence, J. B. Guerard & G. R. Reeves (Eds), *Advances in mathematical programming and financial planning*, 1 (pp. 139–151). Greenwich, CT: Jai Press.

Garstka, S. (1977). Models for computing upper error limits in dollar-unit sampling. *Journal of Accounting Research*, *38*(4), 407–415.

Gary, M. R., & Johnson, D. S. (1979). *Computers and intractability: A guide to the theory of NP-completeness*. New York: W. H. Freeman.

Giokas, D., & Vassiloglou, M. (1991). A goal programming model for bank assets and liabilities management. *European Journal of Operational Research*, *5*(1), 48–60.

Goedhart, M., & Spronk, J. (1995). Interactive decentralized planning: Some numerical experiments. In: H. G. Tzeng, H. F. Fang, U. P. Wen & P. L. Yu (Eds), *Mfultiple criteria decision making, proceedings of the tenth international conference: Expand and enrich the domains of thinking and application* (pp. 265–274). New York: Verlag.

Glover, E., & Greenberg, H. (1989). New approaches for heuristic search: A bilateral linkage with artificial intelligence. *European Journal of Operations Research*, *39*, 119–130.

Goldberg, D. (1989). *Genetic algorithms in search optimization and machine learning*. Reading, MA: Addison-Wesley.

Gonzalez, J. J., Reeves, G. R., & Franz, S. (1987). Capital budgeting decision making: An interactive multiple objective linear integer programming search procedure. In: K. D. Lawrence, J. B. Guerard Jr. & G. R. Reeves (Eds), *Advances in mathematical programming and financial planning*, (Vol. 1, pp. 21–44). Greenwich, CT: Jai Press.

Guerard, J. B., Jr., & Stone, B. K. (1987). Strategic planning and the investment – Financing behavior of major industrial companies. *Journal of the Operational Research Society*, *38*(11), 1039–1050.

Hall, T., Hunton, J., & Pierce, B. (2000). The use of and selection biases associated with nonstatistical sampling in auditing. *Behavioral Research in Accounting*, *12*, 231–255.

Hall, T., Hunton, J., & Pierce, B. (2002). Sampling practices of auditors in public accounting, industry, and government. *Accounting Horizons*, *16*(20), 125–136.

Hitzig, N. (1998). Detecting and estimating misstatement in a two-step sequential sampling with probability proportional to size. *Auditing: A Journal of Practice and Theory*, *17*(1), 54–68.

Holland, J. A. (1975). *Adaptation in natural and artificial systems*. Ann Arbor: University of Michigan Press.

Horgan, J. (1997). Stabilizing the sieve sample size using PPs. *Auditing: A Journal of Practice and Theory*, *62*(2), 40–51.

Johnson, D. S., Aragon, C. R., McGeoch, L. A., & Schevon, C. (1989). Optimization by simulated annealing: An experimental evaluation, Part I, graph partitioning. *Operations Research*, *Vol. 37*, 865–892.

Keown, A. J., & Martin, J. D. (1976). An integer goal programming model for capital budgeting in hospitals. *Financial Management*, *5*(3), 28–35.

Kraft, W. (1968). Statistical sampling for auditors: A new look. *Journal of Accountancy.*, *19*(4), 376–390.

Lawrence, K. D., & Marose, R. A. (1993). Multi-decision-maker, multicriteria strategic planning for the mutual life insurance company. In: K. D. Lawrence, J. B. H. Guerard Jr. & G. R. Reeves (Eds), *Advances in mathematical programming and financial planning*, (Vol. 3, pp. 271–295). Greenwich, Connecticut: Jai Press.

Loebbeck, J., & Neter, J. (1975). Consideration in choosing statistical sampling procedures in auditing. *Journal of Accounting Research*, *13*(Supplement).

Lundy, M., & Mees, A. (1986). Convergence of annealing algorithm. *Mathematical Programming*, *34*, 111–124.

McCray, J. (1973). Ratio and difference estimation in auditing. *Management Accounting*, *55*(December).

Neter, J., Leitch, R., & Fienberg, S. (1978). Dollar unit sampling: Multinomial bounds for total overstatement and understatement errors. *Accounting Review*, *32*(1), 75–81.

Sheshai, K. M. El, Harwood, G. B, & Hermanson, R. H. (1977). Cost volume profit analysis with integer goal programming. *Management Accounting*, *LIX*(October), 43–47.

Rashid, M., & Tabucanon, M. T. (1991). An integrated multi-criteria approach for selecting priority industries for investment promotion. *Information and Management Sciences*, *2*(2), 73–90.

Teitlebaum, A., & Robinson, C. (1975). The real risk in audit sampling. *Journal of Accounting Research*, *13*(Supplement).

Teitlebaum, L., & Schwartz, M. (1958). Practical improvements in audit testing. *Internal Auditor*, *126*(9), 448–453.

A MULTI-CRITERIA FIXED CHARGE PROBLEM: LOCATION OF SERVICE CENTERS

Kenneth D. Lawrence, Sheila M. Lawrence, Ronald Klimberg and Jerry Fjermestad

ABSTRACT

Frequently, problems involving a management activity involve the incurring of a setup charge for the activity. In these cases, the total cost of the activity is the sum of the variable cost related to the level of the activity and a set up cost required to initiate the activity. In this research, we will also use goals that involve demand management of service centers based upon the revenue potential of the customer districts they will serve.

INTRODUCTION

The fixed charge problem is a variant of the linear programming problem where the objective involves a fixed charge. For example, a non-negative decision variable, x_j may have the following cost function:

$$F(x) = d + cx_j \quad x_j > 0 \tag{1}$$

$$0 \quad x_j = 0 \tag{2}$$

Applications of Management Science: In Productivity, Finance, and Operations
Applications of Management Science, Volume 12, 77–88

ISSN: 0276-8976/doi:10.1016/S0276-8976(06)12007-6

That is, there is a fixed initial or set-up cost, d, which must be paid before the continuous decision variable, x, can be used at any non-zero level. Thereupon, the usual variable cost, c, of linear programming, applies. If such fixed charges are non-negative, which is typically true, then they can be modeled in mixed integer programs by using new fixed charge variables. The formulation of the single objective fixed charge problem is as follows:

$$\text{Min } Z = c_j\, x_j + d_j\, y_j \tag{3}$$

$$x_j \geq 0 \tag{4}$$

$$y_j = 0, 1 \tag{5}$$

$$x_j\,(1 - y_j) \geq 0 \tag{6}$$

y_j is an indicator variable of whether or not activity j is undertaken. The set of constraints guarantee that $x_j \geq 0$ only if a corresponding binary variable $y_j = 1$. These are sometimes referred to as a switching constraint.

A particular fixed charge problem is that of the warehouse location problem. These are n customers; the jth one requires b_j units of a particular commodity. There are m locations in which plants may operate to satisfy the demands. There is a fixed charge of d_i for operating plant i. The cost per unit of shipping customer j from plant i is C_{ij} . The capacity of plant i is h_j (Beroggi & Wallace, 1995; Chan, 2000; Chen, Hansen, & Hannard, 1998; Haghan, 1991; Hansen, Mladenovic & Taillard, 1998; Hedge & Tadikmalla, 1990; Mirchardani & Reilly, 1987). The problem formulation is as follows:

$$\text{Min } Z = \sum_{i=1}^{m} \left(\sum_{j=1}^{n} c_{ij} x_j + d_i y_j \right) \tag{7}$$

$$\text{s. t.} \sum_{i=1}^{m} x_{ij} = b_j, \quad j = 1, 2, \ldots, n \tag{8}$$

$$\sum_{j=1}^{n} x_{ij} - h_i y_i \leq 0, \quad i = 1, 2, \ldots, m \tag{9}$$

$$x_{ij} \geq 0 \;\; y_i = 0, 1, \quad \text{for all } i \text{ and } j$$

A MULTICRITERIA FIXED CHARGE PROBLEM AND ITS APPLICATIONS TO SERVICE CENTERS

Decision Variables

This paper will focus on an extension of the fixed charge problem that deals with the warehouse location. In this particular case, several service centers will be located to serve customers in various districts. Service centers provide on-premises service to both business and residential customers in such areas as office equipment, appliances, and computers.

x_{ij} is the number of trucks sent from service center i to district j, where

$$i = 1, 2, 3, \dots, m$$

$$j = 1, 2, 3, \dots, n$$

$$y_i = \begin{cases} 1 & \text{if the service center is leased} \\ 0 & \text{otherwise} \end{cases}$$

GOAL CONSTRAINTS

Total Cost Goal

This minimizes the over-attainment of the total cost goal.

$$\sum_{j=1}^{n}\sum_{i=1}^{m} c_{ij}x_{ij} + \sum_{j=1}^{n} L_i y_i + d^{-}_{\mathrm{TCG}} - d^{+}_{\mathrm{TCG}} = \mathrm{TC_G} \tag{10}$$

L_i is the leasing cost per month for service center, where $i = 1, 2, 3, 4, \dots, m$; c_{ij} the cost per month per truck from service center i to district j; TC_G the goal level of total cost; d^{-}_{TCG} the under-attainment of goal level of total cost, and d^{+}_{TCG} the over-attainment of goal level of total cost.

Demand Satisfaction Goal jth Customer Districts

This is to satisfy demand of the highest priority customer districts.

$$x_{11} + x_{21} + x_{31} + \cdots + x_{m1} + d^{-}_{\mathrm{D1}} - d^{+}_{\mathrm{D1}} = D_1 \tag{11}$$

$$x_{12} + x_{22} + x_{32} + \cdots + x_{m2} + d^{-}_{\mathrm{D2}} - d^{+}_{\mathrm{D2}} = D_2 \tag{12}$$

$$\vdots$$

$$x_{1n} + x_{2n} + x_{3n} + \cdots + x_{mn} + d^{-}_{Dn} - d^{+}_{Dn} = D_n \tag{13}$$

D_j is the demand at the jth customer district, where $j = 1,2,3, \dots n$

Objective Function of the Goal Programming Model

$$\begin{aligned}\text{Min } Z = {} & P_1 d_{\text{TCG}}^{-} + P_{21}\left(d_{D1}^{-} - d_{D1}^{+}\right) + P_{22}\left(d_{D2}^{-} - d_{D2}^{+}\right) \\ & + P_{23}\left(d_{D3}^{-} - d_{D3}^{+}\right) \cdots + P_{2j}\left(d_{Dj}^{-} - d_{Dj}^{+}\right)\end{aligned} \tag{14}$$

Priority 1: Minimize the over-attainment of the goal level of total cost
Priority 2: Satisfy the demand for service center for the jth customer district

$$(P_{21} > P_{22} > P_{23} > \cdots > P_{2j}) \tag{15}$$

The priorities are based upon the revenue potential of the customer districts. The model will be optimized using pre-emptive priority goal programming with mixed integer decision variables.

Hard Constraints Guarantee that Demand is Stratified

$$x_{11} + x_{12} + x_{13} + \cdots + x_{1n} - T_1 y_1 \leq 0 \tag{16}$$

$$\begin{gathered} x_{21} + x_{22} + x_{23} + \cdots + x_{2n} - T_2 y_2 \leq 0 \\ \vdots \end{gathered} \tag{17}$$

$$x_{j1} + x_{j2} + x_{j3} + \cdots + x_{jn} - T_n y_n \leq 0 \tag{18}$$

T_i is the number of trucks, where $i = 1,2,3,4, \ldots n$
Total number of service centers

$$\sum_{i=1}^{M} y_i \leq S \tag{19}$$

S is the number of service centers

EXAMPLE PROBLEM

Tag Corporation is a major appliance corporation, offering a full line of washers, dryers, dishwashers, refrigerators, and ranges. The regional management of Tag is faced with the problem of establishing its service centers in the now growing customer base area of northern New Jersey. Five areas have been identified as premium customers, namely Newark, Princeton, East Brunswick, Bridgewater, and Secaucus.

In order to provide effective service to each area, service centers have been set up that dispatch trucks to customers. If demand must be satisfied, service centers may only be leased for a fixed cost in addition to the variable cost of trucks. Managers have been tasked with controlling costs and satisfying demand.

The management at Tag has to solve the problem of establishing its service centers in the now growing customer base area of northern New Jersey. In this problem, there are five different customer districts/regions. Each facility has a fixed cost (leasing) and has a finite capacity base to the number of trucks available per facility. Each district/region has a potential demand requirement. There is to be considered traveling costs to provide service to the different districts/regions from each facility. We need to determine the locations to lace facilities with minimum cost, which is the sum of facility costs and traveling costs, while satisfying demand requirements of each district/region. Furthermore, management has identified the potential revenue to be expected for serving each district/region; therefore priorities have been set to establish facilities.

The problem now is that the extent of the efficiency depends not only on the location over the facilities but on the scenarios of demand and/or travel costs that are incurred over time.

As mentioned previously, two different objectives are proposed to obtain the final locations. The first one consists of the minimization of the fixed cost per facility. The second one uses Preemptive Goal Programming to prioritize the facilities to be established.

Some of the key features that have been identified of the problem are:

- The management at Tag has multiple goals to solve, whereby it can suggest an ordinal sequence of the order of importance of each of these goals. The problem does not come across as a single goal maximization/minimization problem (minimize the over-attainment of the cost and satisfy customer demand by location, whereby minimizing the under-attainment of the demand).
- Each district further has a relative importance that dictates which district's demand must be serviced first (a particular customer districts' demand must be met because it is higher in priority).

- Data about the fixed cost (required to set up a service center facility and its operational cost is available, as the cost per service trip. The total cost incurred at a location comprises of the fixed leasing cost, plus the cost of service trips (this is given by the product of the total number of trucks shipped with the cost of each shipment made).

The above problem lends itself easily to a goal-programming model. There are various factors, which suggest that a goal-programming approach will be more appropriate. They are:

- multiplicity of goals/objectives and
- ability to define an ordinal hierarchy of priorities for each of the goals that define the order in which they need to be achieved.

Additionally, the presence of both a fixed, as well as variable cost, allows the formulation of the problem as a Fixed Charge problem. A fixed cost is incurred if the management decides to lease the service station at a particular district. The variable cost is a function of the number of trucks shipped. Important to note is the fact that if a service center is not leased to a particular location, there is no variable cost that can be incurred at that location. These factors substantiate the establishment of the model as a Fixed Charge problem.

Thus, the proposed LP model is a combination of the fixed charge problem and goal programming. The model builds on the premise of having a fixed setup/monthly overhead cost of running the facility and the variable cost of servicing a trip/demand request.

The five areas identified as premium customer districts/regions are the following:

1. Newark
2. Princeton
3. East Brunswick
4. Bridgewater
5. Secaucus

These service centers have been identified as high revenue generating locations, and the management is now considering leasing service centers at each of these locations. They are intended to provide on-premise services to both business and residential customers in such as washing machines, dishwashers, refrigerators, etc.

The fixed cost of leasing a service center at each of these locations, as well as the trucks available, is as follows:

Customer District	Leasing Cost ($ 000s)	Trucks Available
Newark	10	12
Princeton	30	13
East Brunswick	23	13
Bridgewater	17	18
Secaucus	12	10

At any given time, the total cost of shipping to a customer is composed of the pro-rata leasing cost and the cost of servicing that customer district.

The table below gives details of the per trip cost for each of these locations

	Customer District (Base) ($)				
Service Center Location	Newark	Princeton	East Brunswick	Bridgewater	Secaucus
Newark	100	786	626	375	300
Princeton	786	110	330	475	505
East Brunswick	626	330	115	425	500
Bridgewater	375	475	425	100	440
Secaucus	300	505	500	440	105

For any cell, the value contained suggests the cost (in dollars) to make a trip from the corresponding district through the row to the corresponding c\district through the column. For example, the cell 23 suggests that the cost of making a trip from Princeton to East Brunswick is $330.

Furthermore, each of these locations has varying demands. The table below gives details on the average demand for each of these districts. The set of average demand has been drawn out from historic data of the past year, and, therefore, serves good estimates of the trends at each location.

Customer District (Base)				
Newark	Princeton	East Brunswick	Bridgewater	Secaucus
14	11	12	14	14

The goals before the management are to decide upon the most profitable service center locations and establish leased servicing facilities there. Once a service center is established at each of these locations, the management moreover needs to reduce the cost of servicing a particular demand. The cost of servicing a demand comes from the combination of the above mentioned fixed leasing cost, and the variable cost per trip. Also, a definite goal before the management is to try to service all the service centers' demands. However, because of revenue considerations, at any given time, the management can establish priorities among these service centers. At any given time, it is more important to serve the demand of a customer district with a higher assigned priority. This means that the customer district with priority A will be served before one with priority B, if $A > B$.

These priorities are given in the following table:

Customer District (Base)				
Newark	Princeton	East Brunswick	Bridgewater	Secaucus
2	1	3	5	4

Where 5 is the highest priority and 1 is the lowest priority.

Converting into a LP Problem

To convert this to a management science problem, we have added a binary variable, Least/Not Leased to indicate whether we are leasing a particular service center or not. We multiply the total number of available trucks at a service location wit this variable. Thus, if a service center is not leased, then the available trucks at that location become 0, while if it is leased, the available trucks remain equal to that provided by the data table.

When a service center is leased, a total of x_{ij} units denote the number of units shipped from a service center i to a customer district j. This assigned is shown in the following table:

	Customer District (Base)					Lease?	Supply	**Relation**	Trucks Available	Actual Available
Service Center Location	Newark	Princeton	East Brunswick	Bridgewater	Secaucus					
Newark	6	7	2	6	6	1	27	≤	28	28
Princeton	3	0	0	0	0	0	3	≤	25	0
East Brunswick	8	12	6	2	4	1	32	≤	35	35
Bridgewater	1	7	4	2	6	1	20	≤	20	20
Secaucus	7	13	0	2	9	1	31	≤	32	32
Average Demand	25	39	12	12	25					
Relation	=	=	=	=	=					
Demand Satisfied	25	39	12	1	25					

The total cost of making an assignment is denoted as follows:

x_{ij} is the number of trucks sent from service center i to customer district j; L_i the cost of leasing a service center i; $Y_i = 1$ if the service center is leased; $0 =$ otherwise.

$$\sum_{i=1}^{n}\sum_{i=1}^{m} c_{ij}x_{ij} + \sum_{j=1}^{n}$$

The application of the above formula is shown in the following table:

Service Center Location	L. Cost	Customer District (Base) ($)					
		Newark	Princeton	East Brunswick	Bridgewater	Secaucus	Total
Newark	8,000	626.19	5,498.60	1,252.93	2,253.42	1,803.80	1,1434.94
Princeton	20,000	2,340.15	2.23	5.65	6.70	6.67	2,361.40
East Brunswick	18,000	5,000.47	3,958.70	689.59	846.55	1,995.32	1,249.63
Bridgewater	13,000	369.31	3,318.28	1,693.60	198.56	2,632.12	8,211.88
Secaucus	10,000	2,036.41	6,566.09	–	879.67	945.15	10,427.33
		10,372.53	19,343.91	4,184.91	4,184.91	7,383.05	44,926.17
						Total goal	300,000
						Over attained goal	–
						Under attained goal	11,730.86
						TCg	300,000

Preemptive

The order of importance of how the facilities should be established requires the usage of the preemptive goal programming. It can be viewed as weighted goal programming where the first goal is infinitely more important than the second goal, the second goal is infinitely more important than the third goal, etc.

Goals

The management of Tag has decided that its Total Monthly Cost of operation should not exceed $100,000. From historical data, occurrences of under-attainment (i.e., when the total cost of the goal is not met or total cost is less than the budgeted cost) and over-attainment of cost (i.e., when total cost exceeds the budgeted cost) are available.

Since over-attainment of cost means that the company is spending more than the budgeted amount, it would like to minimize this over-attainment. Converting this problem to a goal-programming model, we will have the following details:

Let,c_{ij} be the cost per month per truck from service center i to district j; TCG the goal level of total cost; d^{-}_{TCG} the under-attainment of goal level of total cost; d^{+}_{TCG} the over-attainment of goal level of total cost

The objective function now is to obtain the total cost goal.

The formula for the overall Total Cost, including the under-attainment and the over-attainment factors, becomes:

$$\sum_{i=1}^{n}\sum_{j=1}^{m} c_{ij}x_{ij} + \sum^{n} L_i Y_i + d^{-}_{\mathrm{TCG}} + d^{+}_{\mathrm{TCG}}, \quad j = 1$$

CONCLUSION

Facility location problems have been known to be costly, irreversible and have a wide, long-term impact. The proposed model, however, serves as an efficient base tool for managers to decide which service locations to lease and which not to lease. Also, the model provides an optimum number of trips that a service location should make to as to satisfy the multiple goal constraints.

The proposed model builds on past historical data of up to one year and uses the average values representing this historical performance as the data to be used, while establishing the constraints and objectives mathematically. The model proposes that all five service centers be leased and that the assignment from one service center to the other be as under:

	Newark	Princeton	East Brunswick	Bridgewater	Secaucus
Newark	11	0	0	0	2
Princeton	0	10	0	0	1
East Brunswick	0	0	12	0	0
Bridgewater	3	1	0	13	1
Secaucus	0	0	0	1	10

The above assignment satisfies the demands in order of priority in such a way that the cost function is optimized. We obtain the following cost values:

	Newark	Princeton	East Brunswick	Bridgewater	Secaucus	Total
Newark	$10,000	$1,100.00	$78.60	$16.28	–	$600.00
Princeton	$30,000	$71.94	$1,094.91	–	–	$458.92
East Brunswick	$23,000	$57.95	–	$138.00	–	–
Bridgewater	$17,000	$1,159.32	$458.31	–	$1,300	$440.00
Secaucus	$12,000	–	–	–	$433.33	$1,050.00

The above assignment gives a total cost of $101,667 and, therefore an over-attainment of $4,667.

REFERENCES

Beroggi, E. G., & Wallace, W. A. (1995). Operational control of the transportation of hazardous material: An assessment if alternative decision models. *Management Science*, *41*, 1962–1967.

Chan, Y. (2000). *Location, transport and land use: Modeling spatial-temporal information*. New York: Springer.

Chen, P. C., Hansen, P., & Hannard, T. H. (1998). Solution of the multi-source Weber and conditional Weber problems by D-C programming. *Operations Research*, *46*(4), 548–562.

Haghan, A. E. (1991). Multi-criteria decision making in location. *Modeling Transportation Research Record*, *1328*, 88–91.

Hansen, P., Mladenovic, N., & Taillard, E. (1998). Heuristic solution of the multi-source Weber problem: A p-median problem. *Operations Research Letter*, *22*, 55–62.

Hedge, G. G., & Tadikmalla, P. R. (1990). Site selection for a sure service terminal. *European Journal of Operational Research*, *48*, 77–80.

Mirchardani, P., & Reilly, J. (1987). Spatial distribution design for the firefighting units. In: A. Ghosh & G. Rushton (Eds), *Spatial analysis and location-allocation models* (pp. 186–223). New York: Van Nostrand Reinhold.

IDENTIFYING COMPOUNDS FOR PHARMACEUTICAL COMPANIES' DEVELOPMENT TRACKS USING A MULTI-CRITERIA DECISION MODEL

Karen M. Hogan, Gerard T. Olson and George P. Sillup

ABSTRACT

Pharmaceutical companies are faced with identifying development compounds for their Drug Development Processes (DDPs) that will not only gain approval for sale by the regulatory agencies, such as the Food and Drug Administration (FDA), but also establish a sustainable and profitable market presence. This identification of compounds for the DDP includes projection of objective criteria, such as ability to generate revenue and profitability (Financial) and safety and efficacy (Clinical), as well as more subjective criteria, such as determination of insurance coverage by payers, such as the Centers for Medicare and Medicaid Services and pricing (Reimbursement), ability to produce a product of consistent quality (Manufacturing), and attain approval for sale in a timely manner (Registration). The Analytical Hierarchy Process (AHP) is a

Applications of Management Science: In Productivity, Finance, and Operations
Applications of Management Science, Volume 12, 89–108

ISSN: 0276-8976/doi:10.1016/S0276-8976(06)12008-8

multi-criteria decision model that can integrate both objective and subjective information. This study applies the AHP methodology to the identification of compounds resulting in a dynamic application of the model that can be used by pharmaceutical companies to determine the best compounds to put in the DDP, at a time when the cost of conducting clinical evaluations for development compounds is very high and global market conditions are evolving.

1. INTRODUCTION

The prediction of which compound will become the next blockbuster drug in the pharmaceutical industry is a question that is difficult to answer. According to a 1995 study from Tufts University, for every 10,000 or more compounds screened, only 250 of these compounds will enter into pre-clinical testing. Pre-clinical testing is the stage of drug development where the compounds are used in laboratory and animal testing. Of the 250 compounds that enter pre-clinical testing, only 10 compounds succeed in making it to the testing stages, where patient volunteers take the compounds through different clinical phases of study. These phases, usually referenced as phases 1–3 clinical trials, analyze the compounds for safety, dosage, efficacy, side effects, and adverse reactions to long-term use. On average, of the 10 compounds that are worthy of entering the three clinical trial phases, only one will actually succeed in being approved for patient use by the Food & Drug Administration (FDA). The entire process could easily take the pharmaceutical company anywhere between 10 and 15 years and approximately $800 million per approved drug. Given such a long and costly development process, determining which compounds to develop is a pivotal question for the pharmaceutical company to answer.

The identification of the correct compounds to pursue is made more challenging for the pharmaceutical company due to the existence of multiple criteria and the need to integrate both subjective and objective data into the drug development process (DDP). Data is collected from the company's research scientists, clinical investigators and trial sites, healthcare insurers, government and regulatory agencies, such as the FDA, and financial forecasts of the new drug's potential worldwide market share. Once collected, the integration and determination of the relative importance of the different types of data can impact the success or failure of identifying which compound will become the next blockbuster drug. An effective program for identifying compounds should involve a consistent framework for evaluating the next first-in-class patentable drug. This task is especially troublesome because it entails evaluating objective and subjective information. The purpose of this paper is to develop a framework to help pharmaceutical companies evaluate pipeline compounds to minimize the risk of losses associated with chasing compounds that will not make it through the various stages of the approval process, and to optimize the selection of compounds likely to become a new, patentable, first-in-class drug.

A screening model using the Analytical Hierarchy Process (AHP) is developed that facilitates a systematic evaluation of potential first-in-class drugs by incorporating both qualitative and quantitative information into the decision-making process.

2. A REVIEW OF THE LITERATURE

According to Viswanadham and Narahari (2001), research in the area of modeling drug development has primarily been based around optimization of tasks and the benefit of project teams in the drug development process. The authors evaluate and identify ways of reducing drug development times and using efficient scheduling and critical mass-based resource management. They also show that critical mass-based project teams can compress drug lead times further. Macher and Boerner (2005), investigate the question of whether it is most beneficial to perform the tasks of drug development internally versus externally. Their research suggests that pharmaceutical companies encountering problems with greater complexity and requiring more novel development have greater success when these issues are handled internally. However, technology development, which is not the leading edge, realizes performance advantages when outsourced.

Another important research issue that has been addressed recently is evaluating the important steps in the drug development and manufacturing process. Research by DiFeo (2004) reaffirms the link between manufacturing, chemistry and controls, and reproducible quality of the drug used in clinical studies. Product safety is studied through the development process with specific focus on safety found in the early phases of drug development. DiFeo reviews the critical quality attributes of drug substance and development and the relationship to safety and efficacy of the clinical therapy.

Dimasi and Paquette (2004) explore the development of "me-too" or "follow-on", drugs by the pharmaceutical industry. Dimasi and Paquette claim that some have argued that these drugs are duplicative and wasteful, while others argued that with them comes increased competition in the market place, which leads to cheaper prices for consumers. Their research suggests that entry barriers have fallen over time. The majority of "follow-on" drugs that were created in the last decade were in clinical development prior to the approval of the class break through drug. Thus, development races better characterize new drug development.

Given the paucity of research, specifically addressing selection of compounds for inclusion in a drug development pipeline, and the significant economic consequences of that decision, determining which potential compound to develop is an extremely important decision for the pharmaceutical company. To our knowledge, there is no current research which has attempted to identify both the quantitative and qualitative issues associated with a pharmaceutical company's drug development pipeline. In this paper, we will use the AHP to help the pharmaceutical company make more informed decisions regarding the drugs they produce with the limited resources they have to produce them.

3. THE AHP

AHP is a multi-criteria decision support system developed by Saaty (1980, 1990) that allows a decision maker to structure a complex problem in the form of a hierarchy. The first level of the hierarchy is the goal. In the present problem, the goal is to evaluate potential first-in-class patentable drugs effectively. The second level includes the criteria. Relevant criteria needed to evaluate potential first-in-class patentable drugs includes financial data, issues related to the manufacturing or production of the drug, reimbursement or issues associated with insurance coverage for the drug, issues related to the registration process or approval to sell the drug, and clinical or issues related to establishing the safety and efficacy of the drug.

Depending upon the problem, a third level of the hierarchy may be required. The third level specifies a set of sub-criteria related to the criteria of the second level. For example, a set of sub-criteria must be determined to evaluate the issues related to the manufacturing of a potential first-in-class patentable drug. A ratings scale is specified for the criteria and sub-criteria levels. The final level of the hierarchy includes the alternatives to be evaluated. For the present problem, these represent the potential first-in-class patentable drugs, which may be evaluated either individually or as a group.

The AHP model has been recently applied to a multitude of different corporate and non-corporate problems to improve decision making (see for example, Liberatore & Nydick, 1990; Liberatore, Monahan, & Stout, 1993;

Hogan & Olson, 1999, 2004; Ishizaka & Lusti, 2004; Travana, 2004; and Dey, Hariharan, Kumar, & Moseley, 2004). The major advantage of the model is its ability to accommodate complex qualitative and quantitative information into the decision-making process. Other advantages include its simplicity to use and its ability to apply consistency to the decision-making process.

There are four general steps required to implement the AHP. First, the decision maker identifies the criteria and determines their relative importance in achieving the goal, and the sub-criteria and determines their relative importance in achieving the related criterion. Second, the decision maker determines the relative importance of the ratings categories for each of the sub-criteria. Third, the alternatives are evaluated in the context of each of the ratings categories. Finally, the results are synthesized to compute the relative contribution of the alternatives in achieving the goal.

One set of criteria that can be used to evaluate potentially first-in-class patentable drugs includes financial data, issues related to the manufacturing and production of the drug, reimbursement issues, registration issues, and clinical issues related to the development and testing of the drug (other criteria, of course, could also be used). Pairwise comparisons must be made to determine the relative importance of the criteria in achieving the goal. Although there are many scales that can be used to compare the criteria, Saaty (1990) recommends a scale from 1 to 9 where 1 refers to "equally important", 3 "moderately more important", 5 "strongly more important", 7 "very strongly more important", and 9 "extremely more important". If more discrimination is necessary, intermediate values, 2, 4, 6, and 8, can be used. For example, if the pharmaceutical company decides that financial characteristics of the potential drug are "strongly more important" than the registration characteristics of the drug, a value of 5 is assigned to this comparison.

The results of the comparisons are represented in a pairwise comparison matrix similar

$$= \begin{pmatrix} w_{11} & w_{12} & \dots & w_{1n} \\ w_{21} & \dots & \dots & w_{2n} \\ \dots & \dots & \dots & \dots \\ w_{n1} & w_{11} & \dots & w_{nm} \end{pmatrix} = \begin{pmatrix} 1 & w_{12} & \dots & w_{1n} \\ 1/w_{12} & 1 & \dots & w_{2n} \\ \dots & \dots & 1 & \dots \\ 1/w_{2n} & 1/w_{2n} & \dots & 1 \end{pmatrix} \tag{1}$$

where w_{ij} is the relative importance of criteria i compared to criteria j; $w_{ij} = 1 \ \forall \ i = j$; and $w_{ji} = 1/w_{ij} \ \forall \ i \neq j$.

If $n = 5$, W will be a 5×5 matrix with 1's along the main diagonal depicting comparison of the criteria with itself. Below the main diagonal are the reciprocals of the corresponding comparisons above the diagonal. Thus, in this example a total of 10 comparisons must be made. In general, if there are n criteria to be compared, a total of $n(n–1)/2$ comparisons are required.

4. AN APPLICATION OF AHP TO THE IDENTIFICATION OF POTENTIAL FIRST-IN-CLASS PATENTABLE DRUGS

The application of AHP to the identification of potential first-in-class patentable drugs involves four specific steps:

1. Identify the set of criteria and sub-criteria that will be used to evaluate potential first-in-class patentable drugs.
2. Conduct pairwise comparisons of the relative importance of the criteria (and sub-criteria) in evaluating potential first-in-class patentable drugs, and compute criteria (and sub-criteria) weights based on this information.
3. Obtain data that assess the extent a potential new drug achieves the criteria (and sub-criteria).
4. Compute the total assessed weights of the potential patented drug in achieving the goal.

Suppose that a pharmaceutical company wishes to identify which of the numerous compounds they have in their discovery department, or can acquire from academia through licensing agreements is most likely to become a first-in-class patentable drug. The cost of the average drug reaching approval for sale has been increasing dramatically over the last 30 years. Dimasi, Hansen, and Grabowski (2003) find the average discovery and development cost, indexed to year 2000s dollars, to establish the clinical relevancy of a compound so that it can be marketed as a new drug to be \$138 million in the 1970s, \$318 million in the 1980s, and \$802 million in the 1990s. Furthermore, it takes an average of 10 years to attain approval. With such a heavy burden on Research & Development (R&D), it becomes critical to the firm's long-run survival to select compounds that will make it successful through the drug pipeline and emerge as a product approved for sale by regulatory agencies, such as the FDA.

Through interviews with individuals working in the various phases of the DDP in the pharmaceutical industry, we have identified five criteria that can be used to identify potential first-in-class patentable drugs. The five criteria are the financial implications of the development and distribution of the patented drug (Financial), how efficiently the drug can be made (Manufacturing), factors affecting the amount insurers are willing to pay for the drug (Reimbursement), what data it will take to gain registration from regulatory agencies to acquire approval-for-sale status for the drug (Registration), and the safety and efficacy or clinical relevance of the drug (Clinical).

To understand how these five criteria are integrated into the DDP, the activities that occur in the DDP are described in a theoretical timeline as follows:

- phase 0 – a compound is determined to have the capability to mediate disease;
- phase 1 – a compound is determined to be "safe" or cause no harm to patients;
- phase 2 – a compound is determined to be "effective" in treating a patient;
- phase 3 – a compound is determined to be safe and efficacious at a clinically and statistically significant level under the constraints of a clinical protocol;
    - o Product Launch – following approval for sale as a new drug by regulatory agencies in global markets;
- phase 4 – a new drug is prescribed to treat patients in routine practice and monitored for clinical improvement and/or side effects, and for coverage by insurers.

The five criteria can be plotted as milestones on a theoretical DDP timeline of a potential first-in-class patentable compound, so that they can be viewed in the context of the DDP. From the time a compound is identified to go into the DDP, activities to meet the five criteria need to occur at time intervals that will ensure that the new drug will be a financial success, capable of being manufactured, and generate convincing clinical data necessary to attain registration and reimbursement status. Note that the five criteria can be represented more than once so that they coincide with the appropriate phase in the DDP. For example, Clinical is in phase 1 when the drug is first used to treat patients and used again to identify the start of pivotal clinical studies. Given the time and investment to achieve these milestones, it becomes apparent that a multi-criteria decision support system, such as AHP, would be useful in determining which compounds are entered into the DDP.

The first step in the process is to construct a pairwise comparison matrix. In Panel A of Table 1, a pairwise comparison matrix is presented for the potential "innovative" first-in-class patentable drug. If the pharmaceutical company believes that Financial characteristics are "moderately more important" than the Manufacturing characteristics of the first-in-class patentable drug, a value of 3 is placed in cell w_{12} and 0.333 is placed in cell w_{21}. If Financial characteristics are "strongly more important" than the Reimbursement and Registration criteria, a value of 5 is placed in cells w_{13} and w_{14} respectively and 0.2 in cells w_{31} and w_{41} respectively. If Financial is "equally to moderately more important" than Clinical, then a value of 2 is placed in cell w_{15} and a value of 0.5 is placed in cell w_{51}. Other pairwise comparisons for the five criteria are also presented in Table 1.

The pairwise comparison matrix can be used to estimate the criteria weights. To determine the criteria weights, each column of the matrix in Panel A is summed, and each cell is divided by its column total. The result is

Table 1. Pairwise Comparison Matrix and Computations for Compound Evaluation Criteria.

Panel A: Pairwise Comparison Matrix

	Financial	Manufacturing	Reimbursement	Registration	Clinical
Financial	1.0000	3.0000	5.0000	5.0000	2.0000
Manufacturing	0.3333	1.0000	4.0000	3.0000	0.3333
Reimbursement	0.2000	0.2500	1.0000	0.5000	0.2500
Registration	0.2000	0.3333	2.0000	1.0000	0.2000
Clinical	0.5000	3.0000	4.0000	5.0000	1.0000
Column total	2.2333	7.5833	16.0000	14.5000	3.7833

Panel B: Adjusted Comparison Matrix and Criteria Weights

	Financial	Manufacturing	Reimbursement	Registration	Clinical	Criteria weights
Financial	0.4478	0.3956	0.3125	0.3448	0.5286	0.4059
Manufacturing	0.1493	0.1319	0.2500	0.2069	0.0881	0.1652
Reimbursement	0.0896	0.0330	0.0625	0.0345	0.0661	0.0571
Registration	0.0896	0.0440	0.1250	0.0690	0.0529	0.0761
Clinical	0.2239	0.3956	0.2500	0.3448	0.2643	0.2957
					Total	1.0000

the adjusted comparison matrix that is presented in Panel B of Table 1. The criteria weights are calculated by computing the average of the entries in each row. In this example, the final weights for Financial, Manufacturing, Reimbursement, Registration, and Clinical criteria are 0.4059, 0.1652, 0.0571, 0.0761, and 0.2957 respectively. This can be interpreted as Financial criteria being 2.457 times more important than Manufacturing criteria (0.4059/0.1652), 7.096 times more important than Reimbursement (0.4059/0.1652), 5.325 times more important than Registration (0.4059/0.0761), and 1.370 times more important than Clinical (0.4059/0.2957). [1]

Once the criteria weights are computed, the pharmaceutical company can specify a set of sub-criteria and determine their relative importance. Using the approach described above, weights for the sub-criteria can be similarly computed. Alternatively, the decision maker can input the ratings directly. These weights measure the importance of each of the sub-criteria relative to its criterion. Similar to the set of criteria, the choice of appropriate sub-criteria is dependent upon an evaluation of those factors that can best discriminate between appropriate pipeline compounds.

On the basis of research and survey information in the field, Financial characteristics include unit price times number of units sold (Revenues), time to market (Time), Break Even Point, Profit Margin, estimate of costs (Outlay), and net present value (NPV).

Manufacturing characteristics include decisions regarding the plant operation (Plant), equipment used to produce the drug (Equipment), work process specification (Process), and certification regarding good manufacturing practices (Certification).

Reimbursement characteristics will include decisions regarding insurance coverage at launch of the drug (Coverage), coding for reimbursement (e.g., ICD-10 codes for insurance reimbursement) at launch (Coding), pricing of the product at launch (Pricing), and cost justification for the price of the drug (Pharmacoeconomics).

Registration characteristics look at variables, such as analyzing the clinical data (Analysis), developing a clinical summary of the data (Summary), submitting the data and analysis for registration of the drug (Submission), and meetings with regulatory agencies, e.g., the FDA, about the drug's registration (FDA Meetings).

Clinical characteristics involve looking at the analyzability of the clinical data collected (Analyzable Data), time and effort to collect the data (Data Collection), institutional review board approval for the clinical trial (Approval), the investigator/site approval for the clinical trial (Site Selection), and the clinical study plan (Protocol).

Table 2. Compound Evaluation Criteria and Sub-Criteria Local Weights.

Financial .4059		Manufacturing .1652		Reimbursement .0571		Registration .0761		Clinical .2957	
Revenue .0522	> $1B .8 $.5B - $1B .5 < $.5B -.3	Plant .3801	Existing .8 Renovate .4 New -.2	Coverage .5328	Launch 1.5 < 6 months -.2 > 6 mo -.3	Analysis .0418	< 3 months .8 3 -6 months .5 > 6 mo -.3	Analyzable Data .4782	< 3 months .8 3 - 6 months .4 > 6 mo -.2
Time .4141	Phase 2 .5 Fast Track .3 Phase 3 .2	Equipment .2219	Existing .8 Some New .4 All New -.2	Coding .0559	Launch 1.5 < 6 mo -.1 > 6 months -.4	Summary .0834	< 3 mo .8 3 - 6 months .5 > 6 months -.3	Data Collection .2672	Faster .8 On Time .4 Later -.2
Break Even Point .1637	< 6 mo .5 6 – 18 months .4 > 18 mo .1	Process .1971	Revalidate .8 Revise .4 Write -.2	Pricing .2858	Launch 1.8 After -.8	Submission .2463	< 3 mo .8 3 - 6 months .5 > 6 mo -.3	Approval .1131	< 3 mo .8 3 - 6 months .3 > 6 mo -.1
Profit Margin .0396	> 30% .6 20 - 30% .3 < 20% .1	Certification .2008	< 6 mo .7 6 - 12 months .2 > 12 months .1	Pharmaeconomics .1255	>12mo 1.5 6 - 12 months -.2 < 6 months -.3	FDA Meetings .6285	< 6 mo .6 6 - 12 months .3 > 12 months -.1	Site Selection .0975	<3 mo .8 3 - 6 months .3 > 6 mo -.1
Outlay .2205	<$400m .6 $400-$800m .3 >$800m .1							Protocol .0440	< 6 mo .7 6 - 12 months .2 >12mo .1
	>$200m 1.2								

Table 2 presents the criteria and sub-criteria local weights for an innovative first-in-class patentable drug. The local weight for item *i* refers to the proportion of value placed on item *i* relative to its parent. The weights for the sub-criteria can be determined in the same fashion as those described above. Also listed in Table 2 are ratings scales and their weights for the sub-criteria.

Suppose the pharmaceutical company determines a compound's Break Even Point, it provides more information concerning Financial characteristics than NPV, Revenue, and Profit Margin, but less information than Outlay and Time. This can be incorporated into the weights in Table 2 of the sub-criteria affecting Financial characteristics. In particular, Break Even Point is viewed by the pharmaceutical company as "moderately more important" than Revenue and NPV. Break Even Point is viewed as "moderately to strongly more important" than Profit Margin. Time is determined to be "moderately more important" than Break Even Point. In addition, Time is viewed as "very strongly more important" than Revenue, "extremely more important" than Profit Margin, "strongly more important" than NPV, and "equally to moderately more important" than Outlay. Revenue is "equally to moderately more important" than Profit Margin. Outlay is "strongly more important" than Revenue, "equally to moderately more important" than Break Even Point and NPV, and "moderately to strongly more important" than Profit Margin. NPV is "moderately more important" than Profit Margin and "moderately to strongly more important" than Revenue. Using this information, the sub-criterion weights are computed as 0.0522, 0.4141, 0.1637, 0.0396, 0.2205, and 0.1100 for Revenue, Time, Break Even Point, Profit Margin, Outlay, and NPV respectively.

The Revenue sub-criteria are rated based on a three-point dollar scale. If the Revenue is less than $500 million, then a negative weight of 0.3 is assigned. If the Revenue falls between $500 million and $1 billion, then a weight of 0.5 is used. For Revenue greater than $1 billion, a weight of 0.8 is assigned.

The Time sub-criteria are rated based on a three-point scale. If the Time variable has Phase 2 favorable results, then a weight of 0.5 is assigned. If Time has the drug on a fast-track then a weight of 0.3 is used. For Time that has the drug with pivotal clinical studies in Phase 3, a weight of 0.2 is assigned.

The Break Even Point sub-criteria are also rated based on a three-point scale. If the Break Even Point is six months or less after launch of the drug then a weight of 0.5 is assigned. If the Break Even Point falls between six months and 18 months, a weight of 0.4 is used. For a Break Even Point greater than 18 months, a weight of 0.1 is assigned.

The Profit Margin is denoted with a three-point scale. If Profit Margin is greater than 30%, a weight of 0.6 is used. If Profit Margin falls between 20–30%, then a weight of 0.3 is used. For a Profit Margin less than 20%, a weight of 0.1 is assigned.

Outlay sub-criteria are determined based on a three-point scale. If the Outlay is under $400 million, then 0.6 weight is assigned. If the Outlay falls between $400 and $800 million, then a weight of 0.3 is used, and for Outlay greater than $800 million, a weight of 0.1 is assigned.

Finally, the NPV sub-criteria are evaluated on a three-point dollar scale. If the NPV is greater than $200 million, then a weight of 1.2 is used. If the NPV is between $0 and $200 million, a weight of 0.2 is used, and for a negative NPV, a weight of –0.4 is used.

From prior drug development, the pharmaceutical company determines that when evaluating the Manufacturing aspects of producing the drug, the Plant is ‘‘slightly more important’’ than Equipment, Process, and Certification. This can be incorporated into the weights in Table 2 of the sub-criteria affecting the Manufacturing. In particular, Plant is ‘‘equally to moderately more important’’ than Equipment and Process, but ‘‘moderately more important’’ than Certification. Process is viewed as ‘‘equally to moderately more important’’ than Equipment. Certification is viewed as ‘‘moderately more important’’ than Process. Using this information, the sub-criterion weights are computed as 0.0381, 0.2219, 0.1971, and 0.2008 for Plant, Equipment, Process, and Certification respectively.

The Plant sub-criteria are rated based on a three-point scale. If the Plant will use an existing Good Manufacturing Practices (GMP) facility, then a weight of 0.8 is assigned. If the pharmaceutical company will need to renovate an existing facility to manufacture the product, a weight of 0.4 is used, and if building a new facility is necessary, a weight of –0.2 is assigned.

The Equipment sub-criteria are rated based on a three-point scale. If it is possible to revalidate existing equipment to use in production, then a weight of 0.8 is assigned. If the company can revalidate some equipment, but needs to also purchase some equipment, then a value of 0.4 is assigned and, if all new equipment must be purchased, a weight of –0.2 is used.

The Process specification sub-criteria are rated based on a three-point scale. If the pharmaceutical company can revalidate existing specifications, then a weight of 0.8 is assigned. If the work process entails both revising and revalidating the existing specifications, then a weight of 0.4 is used and, if the company needs to both write and validate the work process specifications, then a value of –0.2 is used.

The Certification sub-criteria are evaluated based on a three-point scale. If the company can get new GMP certification in less than or equal to six months, a value of 0.7 is assigned. If the certification process takes between 6–12 months, then a value of 0.2 is used and, greater than a year, a value of 0.1 is assigned.

From a prior experience with different Reimbursement issues related to a new drug, the pharmaceutical company determines that Coverage is more important than Coding, Pricing, and Pharmacoeconomics. They also determine that Pricing of the drug is more important to the firm than Coding and Pharmacoeconomics. However, Pharmacoeconomics is more important than Coding. This can be incorporated into the weights in Table 2 of the sub-criteria affecting Reimbursement. In particular, Coverage is "strongly more important" than Coding, "moderately more important" than Pricing, and "very strongly more important" than Pharmacoeconomics. Pricing is viewed as "very strongly more important" than Coding and "moderately to strongly more important" than Pharmacoeconomics. Pharmacoeconomics is "strongly more important" than Coding. Using this information, the sub-criterion weights are computed as 0.5328, 0.5559, 0.2858, and 0.1255 for Coverage, Coding, Pricing, and Pharmacoeconomics respectively.

The Coverage sub-criteria are rated based on a three-point scale. If payment levels are established at launch of the drug, then a weight of 1.5 is assigned. If payment levels are established in less than six months after launch, then a weight of −0.2 is used and, if it takes more than six months, then a weight of −0.3 is assigned.

The Coding sub-criteria are rated based on a three-point scale. If the drug is coded for Reimbursement at product launch, a weight of 1.5 is assigned. If the drug is coded within six months after product launch, a weight of −0.1 is used and, if the coding takes longer than six months, a value of −0.4 is used.

Pricing sub-criteria are rated based on a two-point scale. If the pharmaceutical company has pricing of the product at launch to include all discounts, then a value of 1.8 is assigned. If the product's price is unresolved at launch, for example due to unresolved discounts, a value of −0.8 is assigned.

The Pharmacoeconomics sub-criteria are based on a three-point scale. If the Pharmacoeconomics of the drug is available more than a year prior to launch, then a weight of 1.5 is assigned. If the Pharmacoeconomics are available between 6–12 months prior to launch, then a weight of −0.2 is used and, if they are available between 0–6 months after launch, a weight of −0.3 is assigned.

From an analysis of the former drug registrations, the drug company determines that meetings with regulatory agencies required for the approval of the new drug are "extremely more important" than Analysis and Submission. These FDA Meetings are seen as "very strongly more important" than the Clinical Summary. The Submission for registration is "very strongly more important" than both Analysis and Summary is "moderately more important" than Analysis. Using this information, the sub-criterion weights are computed as 0.0418, 0.0834, 0.2463, and 0.6285 for Analysis, Summary, Submission, and FDA Meetings respectively.

The Analysis sub-criteria are rated based on a three-point scale. If the drug company believes that the analysis will take less than three months after receipt of the data, then a weight of 0.8 is assigned. If the analysis will take between three to six months, a weight of 0.5 is used and, if the analysis will take longer than six months, then a weight of −0.3 is used.

The Clinical Summary sub-criteria are rated based on a three-point scale. If the pharmaceutical company will require less than three months for the clinical summary, then a weight of 0.8 is assigned. If the company believes the clinical summary may take between three to six months, then a value of 0.5 is used and, if greater than six months, a weight of −0.3 will be assigned.

The Submission for registration sub-criteria are also rated based on a three-point scale. If the drug company is able to submit for registration less than three months after the analysis and clinical summary, then a value of 0.8 is assigned. If the submission will take between three and six months, then a weight of 0.5 is used and, if greater than six months, a value of −0.3 will be required.

Finally, the FDA Meetings sub-criteria are rated based on a three-point scale. If meetings with the FDA are in less than six months after submission for registration, then a value of 0.6 is used. If the meetings are 6–12 months after submission, then a value of 0.3 is used and, if greater than 12 months, a weight of 0.1 is used.

Looking at other Clinical studies the pharmaceutical company has undertaken, they decide that the time it takes for them to have access to an Analyzable Data set is "moderately more important" than the Data Collection, "strongly more important" than the Internal Review Board Approval and Site Selection. In addition, the Analyzable Data set is "very strongly more important" than the clinical Protocol. The Data Collection is "moderately to strongly more important" than Approval and Site Selection, and "strongly more important" than Protocol. Approval is felt to be "equally to moderately more important" than Site Selection, and "moderately more important" than Protocol. Site Selection is shown to be "moderately to strongly more important" than Protocol. Using this information, the sub-criteria weights are determined to be 0.4782, 0.2672, 0.1131, 0.0975, and 0.0440 for Analyzable Data, Data Collection, Approval, Site Selection, and Protocol respectively.

AHP can help the pharmaceutical company focus on the determination of the weights or "priorities" of each criteria and sub-criteria in accomplishing the goal. The weights represent the relative importance the pharmaceutical company places on each of the attributes. Once the weights for each of the attributes are determined, the pharmaceutical company can assess a potential first-in-class patentable drug on each of the attributes, and determine whether an investment in the compound should be made by placing it in the DDP.

Table 3 presents the global weights for the entire DDP hierarchy of our hypothetical first-in-class patentable drugs. The global weight for item *i* refers to the proportion of total value placed on item *i* by the researcher.

Table 3. Compound Evaluation Criteria and Sub-Criteria Global Weights.

Financial		Manufacturing		Reimbursement		Registration		Clinical	
Revenue .0212	> $1B .0169	Plant .0628	Existing .0502	Coverage .0304	Launch .0456	Analysis .0032	< 3 months .0025	Analyzable Data .1414	< 3 months .1131
	$.5B - $1B .0106		Renovate .0251		< 6 months -.0061		3 -6 months .0016		3 – 6 months .0566
	< $.5B -.0064		New -.0126		> 6 mo -.0091		> 6 mo -.0010		> 6 mo -.0283
Time .1681	Phase 2 .0840	Equipment .0367	Existing .0293	Coding .0032	Launch .0048	Summary .0063	< 3 mo .0051	Data Collection .0790	Faster .0632
	Fast Track .0504		Some New .0147		< 6 mo -.0003		3 - 6 months .0032		On Time .0316
	Phase 3 .0336		All New -.0073		> 6 months -.0013		> 6 months -.0019		Later -.0158
Break Even Point .0664	< 6 mo .0332	Process .0326	Revalidate .0261	Pricing .0163	Launch .0294	Submission .0187	< 3 mo .0150	Approval .0334	< 3 mo .0268
	6 – 18 months .0266		Revise .0130		After -.0131		3 - 6 months .0094		3 – 6 months .0100
	> 18 mo .0066		Write -.0065				> 6 mo -.0056		> 6 mo -.0033
Profit Margin .0161	> 30% .0096	Certification .0332	< 6 mo .0232	Pharmaeconomics .0072	>12mo .0107	FDA Meetings .0478	< 6 mo .0287	Site Selection .0288	<3 mo .0231
	20 - 30% .0048		6 - 12 months .0066		6 – 12 months -.0014		6 - 12 months .0143		3 – 6 months .0087
	< 20% .00016		> 12 months .0033		< 6 months -.0021		> 12 months .0048		> 6 mo -.0029
Outlay .0895	<$400m .0537							Protocol .0130	< 6 mo .0091
	$400- $800m .0268								6 – 12 months .0026
	>$800m .0089								>12mo .0013
NPV	>$200m .0536								
	$0- $200m								

Global weights are computed by multiplying the local weight of the item by the local weight of each of its parents. For example, suppose the pharmaceutical company believes Revenue for the new drug will be greater than $1 billion. From Table 2 we observe that the local weight for Revenue is greater than $1 billion is 0.8, the local weight for Revenue is 0.0522, and the local weight for Financial is 0.4059. Thus, the global weight for Revenue greater than $1 billion is 0.0169. Other global weights in Table 3 are computed in a similar fashion.

Once the global weights have been determined, the pharmaceutical company gathers information concerning drug compounds that it is evaluating. Table 4 presents hypothetical data for three compounds that will emerge from the DDP as a potential first-in-class patentable drug. Each of the compounds is rated based on the sub-criteria. For example, the NPV for compound A is expected to be greater than $200 million, while NPV for compounds B and C are expected to be between $0 and $200 million. Thus, using the data from Table 3, compound A earns a score of 0.0536 for NPV and compounds B and C earn a score of 0.0089 for NPV.

The scores for all of the ratings are summed and an overall score is determined. For the potential first-in-class patentable compounds depicted in Table 4, compound A earns a total of 0.3220, compound B 0.4192, and compound C 0.2550. Once the total scores have been computed, a decision can be made whether to invest in any or all of the compounds based on all the current information concerning the compounds. The pharmaceutical company can either choose a maximum cut off for the compound to justify the investment or further segment the scores into categories of high probability of success, further investigation warranted, or low probability of successfully becoming a new drug. For example, using our hypothetical compounds, the pharmaceutical company may want to reject any compounds with scores less than 0.2500, further investigate compounds with scores between 0.2500 and 0.3000, and invest in compounds with scores greater than 0.3000. The result is a structured framework that integrates both qualitative and quantitative information to evaluate all compounds based on the same scale to determine whether they are potential first-in-class patentable drugs.

While these results are excellent references to assess which compounds are entered in the DDP, there are times that one compound is selected over another despite having a higher score in one of the criteria, e.g., NPV. This can be the case even when there are drugs to treat a disease but the drugs have limitations, such as side effects or the inability to treat all patients efficaciously. In this situation, a compound with a slightly lower NPV might

Table 4. Evaluation of Potential Compounds.

	A	Weight	B	Weight	C	Weight
Financial						
Revenue	\$.5–\$1B	0.0106	> \$1B	0.0169	< \$.5B	−0.0064
Time	Phase III	0.0336	Phase III	0.0336	Fast track	0.0504
Break Even Point	> 18 m	0.0066	> 18 m	0.0066	< 6 m year	0.0332
Profit Margin	> 30%	0.0096	< 20%	0.0016	> 30%	0.0096
Outlay	\$400–\$800M	0.0268	> \$800M	0.0089	⩽ \$400M	0.0537
NPV	> \$200M	0.0536	\$0–\$200M	0.0089	\$0–\$200M	0.0089
Manufacturing						
Plant	new	−0.0126	existing	0.0502	new	−0.0126
Equipment	some new	0.0147	All new	−0.0073	all new	−0.0073
Process	revise	0.0130	revise	0.0130	write	−0.0065
Certification	> 12 m	0.0033	6–12 m	0.0066	6–12 m	0.0066
Reimbursement						
Coverage	< 6 m after	−0.0061	at launch	0.0456	at launch	0.0456
Coding	< 6 m after	−0.0003	at launch	0.0048	at launch	0.0048
Pricing	at launch	0.0294	after launch	−0.0131	after launch	−0.0131
Pharmaeconomics	6–12 m	−0.0014	> 12 m before	0.0107	6–12 m	−0.0014
Registration						
Analysis	3–6 m	0.0016	< 3 m	0.0025	> 6 m	-0.0010
Summary	> 6 m	-0.0019	< 3 m	0.0051	3–6 m	0.0032
Submission	> 6 m	-0.0056	3–6 m	0.0094	3–6 m	0.0094
FDA Meetings	6–12 m	0.0143	< 6 m	0.0287	> 12 m	0.0048
Clinical						
Analyzable data	3–6 m	0.0566	< 3 m	0.1131	> 6 m	−0.0283
Data Collection	on time	0.0316	faster	0.0632	faster	0.0632
Approval	< 3 m	0.0268	3–6 m	0.0100	< 3 m	0.0268
Site Selection	3–6 m	0.0087	> 6 m	−0.0029	3–6 m	0.0087
Protocol	< 6 m	0.0091	6–12 m	0.0026	6–12 m	0.0026
Total		0.3220		0.4192		0.2550

be entered into the DDP because it can overcome the limitations of approved drugs.

To illustrate this point, consider that the hypothetical compounds in Table 4 are similar to drugs currently on the market to treat Type 2 diabetes (better known as age-onset). All three compounds have the potential to treat Type 2 diabetes more effectively (better efficacy, fewer side effects, require fewer needle-stick blood tests) than the six oral medications currently on the market. Looking at the Clinical criteria for these compounds, they will be compared to drugs already on the market, which have various advantages to controlling diabetes, but also have significant side effects or expense issues that would make these compounds potentially more desirable. For example, a current drug may be relatively inexpensive but cannot be taken if the patient is allergic to sulfa drugs or a drug that controls blood sugar level, but must be taken in conjunction with meals.

In addition to this Clinical advantage, some of the compounds may be desirable based on the other four criteria. For example, one of the four pharmaceutical companies (Sanofi-Aventis, GlaxoSmithKline, Amyline Pharmaceuticals, Eli Lily) with diabetes products approved for sale and in development, may determine that their Manufacturing capability gives them a decided edge in the market. Given the enormity of the market – about $7.3 billion spent to treat 16 million patients in the U.S. (Diabetes Care, Anonymous, 2003) – a comprehensive assessment of criteria in the AHP model can help a pharmaceutical company make the right choice in placing compounds in the DDP.

5. MODEL VALIDATION

In the above example, the expert knowledge of the pharmaceutical company is used to conduct pairwise comparisons. Since the resulting drug-selection process is dependent upon these comparisons, it is desirable to use the optimal set of criteria and their relative importance. Unfortunately, no researcher has yet identified the optimal set of criteria that can be used to evaluate compounds as potential first-in-class patentable drugs and, depending on the compound, these variables could also change. The model presented here integrates information from both academic research studies and actual drug industry practice to build the best model that can incorporate both quantitative and qualitative data into the DDP to evaluate compounds.

AHP is a multi-criteria decision support system that can integrate quantitative and qualitative information. The value of the model is dependent upon the inputs of the expert. Until an optimal set of criteria can be identified for each possible compound, care should be taken in the selection of the relevant inputs in evaluating potential first-in-class patentable drugs.

For pharmaceutical companies, the model can be used as a focal point for rethinking the tradeoffs among different sets of criteria that are pertinent to the important question of which compounds to place on their DDPs. The simplicity of implementing AHP permits the pharmaceutical company to easily revise the criteria based upon different situations that may arise while evaluating different compounds. Model parameters can be modified until the AHP rankings of analyzed potential compounds are consistent with post drug development information supplied by the pharmaceutical company. Not only will the output of the model be improved but also valuable insight can be gained that may eventually lead to the identification of an optimal set of criteria to evaluate potential first-in-class patentable drugs, and thus save the company from using limited resources to chase the wrong compounds.

6. SUMMARY AND CONCLUSIONS

The effective evaluation of potential first-in-class patentable drugs involves both subjective and objective information. A difficulty arises in the implementation of the process due to multiple evaluative criteria that may be troublesome to measure. AHP is a decision-support system that can integrate both subjective and objective information to improve the efficiency selecting the "best" compound for the pharmaceutical companies to develop.

AHP requires the structuring of the problem into the form of a hierarchy, which consists of a goal, evaluation criteria, and possibly sub-criteria, and alternatives. Pairwise comparisons are made on items on each level of the hierarchy to the level above it and the relative importance of the items is determined. An overall score for each alternative is computed, and can be contrasted to a benchmark established by the decision maker.

In this paper, we described AHP and applied the model to the evaluation of potential first-in-class patentable drugs. The result is a flexible and consistent scoring model that can reduce the risk associated with the pharmaceutical company's investment in the development of potentially patentable drugs. AHP is flexible in that the criteria and sub-criteria can be revised based upon the needs of the user. Also, the relative importance of the criteria and sub-criteria can be easily recomputed with the use of a spreadsheet or dedicated software. The result is a consistent and effective measurement scale that can be used to evaluate the difference between a potential failure and the next big blockbuster drug.

NOTES

1. The procedure presented here is good approximation of the weights. Saaty (1980) determined the exact relative priorities for each of the *n* criteria by computing the normalized eigenvector of the maximum eigenvalue of the comparison matrix. The normalized eigenvector is computed by raising the comparison matrix to successive powers until convergence is achieved, and then normalizing the results.

REFERENCES

Anonymous (2003). Economic costs of diabetes in the US in 2002. *Diabetes Care*, *26*(March), 3.

Dey, P. K., Hariharan, S., Kumar, A., & Moseley, H. (2004). Performance measurement of intensive care services in hospitals: The case of Barbados. *International Journal of Services Technology & Management*, *5*, 1.

DiFeo, T. J. (2004). Safety and efficacy: The role of chemistry, manufacturing, and controls in pharmaceutical drug development. *Drug Development and Industrial Pharmacy*, *30*(3), 247–257.

Dimasi, J. A., & Paquette, C. (2004). The economics of follow-on drug research and development: Trends in entry rates and the timing of development. *Pharmacoeconomics*, *22*(s2), 1–14.

Dimasi, J., Hansen, R., & Grabowski, H. (2003). The piece of innovation: New estimates of drug development costs. *Journal of Health Economics*, *22*, 15.

Hogan, K., & Olson, G. (1999). Evaluating potential acquisitions using the analytic hierarchy process. *Advances in Mathematical Programming and Financial Planning*, *5*, 3–17.

Hogan, K., & Olson, G. (2004). A multi-criteria decision model for portfolio allocation for the individual investor. *Mathematical Programming: Applications of Management Science*, *11*, 3–16.

Ishizaka, A., & Lusti, M. (2004). An expert module to improve the consistency of AHP matrices. *International Transactions in Operational Research*, *11*, 97–105.

Liberatore, M., Monahan, T., & Stout, D. (1993). Strategic capital budgeting for investments in advanced manufacturing technology. *Journal of Financial and Strategic Decision Making*, *6*(Summer), 55–72.

Liberatore, M., & Nydick, R. (1990). An analytic hierarchy approach for evaluating product formulations. In: A. Bohl (Ed.), *Computer aided formulation: A manual for implementation* (pp. 179–196). VCH Publishing Company.

Macher, J. T. & Boerner, C. S. (2005). *Development and the boundaries of the firm: A knowledge-based examination in drug development*. Working Paper, BPS: GG1.

PhRMA, (1995). *Data from Center for the Study of Drug Development*, Tufts University.

Saaty, T. J. (1980). *The analytical hierarchy process*. New York, NY: McGraw-Hill.

Saaty, T. J. (1990). How to make a decision: The analytic hierarchy process. *European Journal of Operational Research*, *48*, 9–26.

Travana, M. (2004). A subjective assessment of alternative mission architectures for the human exploration of Mars at NASA using multicriteria decision making. *Computers & Operational Research*, *31*, 1147.

Viswanadham, N., & Narahari, Y. (2001). Queueing network modeling and lead time compression of pharmaceutical drug development. *International Journal of Product Research*, *39*(2), 395–412.

CALCULATING DISASSEMBLY YIELDS IN A MULTI-CRITERIA DECISION-MAKING ENVIRONMENT FOR A DISASSEMBLY-TO-ORDER SYSTEM

Prasit Imtanavanich and Surendra M. Gupta

ABSTRACT

In this paper, we consider the disassembly-to-order (DTO) problem, where a variety of returned products are disassembled to fulfill the demand for specified numbers of components and materials. The objective is to determine the optimal numbers of returned products to disassemble so as to maximize profit and minimize costs. We model the DTO problem using a multi-criteria decision-making approach. Since the conditions of returned products are unknown, the yields from disassembly are considered to be stochastic. To solve the stochastic problem, we use one of the two heuristic approaches (viz., one-to-one approach or one-to-many approach) that converts the problem into a deterministic equivalent. We compare the performance of the two heuristic approaches using a case example.

Applications of Management Science: In Productivity, Finance, and Operations
Applications of Management Science, Volume 12, 109–125

ISSN: 0276-8976/doi:10.1016/S0276-8976(06)12009-X

INTRODUCTION

In many countries, one of the most important problems of manufacturing companies is to determine how to appropriately manage products at the end of their lives. Due to the rapid development in manufacturing technologies, coupled with the desire of customers to acquire products with the latest technologies, older products become undesirable even though they are still in very good operating conditions. This phenomenon triggers a premature disposal of products leading to shorter life spans, which in turn, generates more waste that impacts the environment in a negative way.

The most common ways to manage the end-of-life (EOL) products include remanufacturing, reuse, recycling and disposal. The least desirable among these is disposal of products. Disposal causes reduction in the number of landfills and is harmful to the environment. This awareness among the governments, manufacturers and consumers has resulted in the establishment of several environmental regulations. As many of these EOL products are in good working conditions, the preferable choices are to reuse, remanufacture or recycle them, in that order. These choices, not only save on the disposal costs, they are actually better for the environment.

Regardless of the choice made to manage the EOL products, the first step in recovering components or materials is disassembly. In this paper, we consider the disassembly-to-order (DTO) problem as a variety of returned products are disassembled to fulfill the demands for specified numbers of components and materials. The objective is to determine the optimal number of returned products to disassemble so as to maximize profit and minimize costs and the environmental detriments. Because the returned EOL products are received in a variety of conditions, it is often difficult to know a priori, exactly how many products would be needed to disassemble to fulfill the demands for components and materials. We capture this behavior using one of the two heuristic approaches (viz., one-to-one approach or one-to-many approach) that convert the problem into an approximate deterministic equivalent.

LITERATURE REVIEW

Many researchers have studied disassembly in recent years. Some of them have focused on the financial aspects of disassembly, while others have studied the disassembly process itself. We briefly review some of the relevant literature.

Moyer and Gupta (1997) presented a survey of works related to environmentally conscious manufacturing, recycling and complexity of disassembly in the electronics industry. Stuart and Qin (2000) proposed models to maximize the net revenue with processing and inventory capacity constraints. Sodhi and Reimer (2001) introduced two mathematical programming models, one to minimize the cost of recycling and the other to maximize profit. Veerakamolmal and Gupta (1998, 1999) presented mathematical models for solving the lot-size balancing problems, which are essentially the DTO problems. These models were based on a combination of linear programming and several nonlinear operations, including the ceil operation (smallest integer larger than the operand). This model was later modified to avoid the ceil operation by Lambert and Gupta (2002). Kongar and Gupta (2002) proposed a multi-criteria decision-making approach for the DTO system. Inderfurth and Langella (2003) introduced one-to-one and one-to-many heuristic approaches for solving the DTO problems under stochastic yields. Imtanavanich and Gupta (2004) extended the one-to-one stochastic yields methodology to handle the DTO problems with multiple objectives. For more information, see Gungor and Gupta (1999) and Lambert and Gupta (2005).

Disassembly is the process of systematic removal of the desired components or materials from the original assembly so that the components or materials are obtained in the desired form (Kongar & Gupta, 2002). In a DTO system, varieties of EOL products are taken back from the last users to the disassembly facility and are disassembled to satisfy the demands for components and materials. In this paper, we consider both non-destructive (focusing on items rather than materials) and destructive (focusing on materials rather than items) disassembly (see Fig. 1).

The DTO process starts with the retrieval of take-back EOL products from collectors. Depending on the condition of the EOL products, some of them are sent to disposal. The rest (majority) of them are disassembled, using either destructive or non-destructive disassembly. The component yields from non-destructive disassembly are stochastic. The components obtained are in either good working condition or broken. Good components are used to satisfy the demands for used components. Additional components may be supplemented from outside suppliers to satisfy the demands. Bad components, together with materials from destructive disassembly, are sent to the recycling process, either in-plant or out of plant, to satisfy the demands for materials used in recycling. In-plant recycling is limited to the available in-plant capacity. Since it is difficult to recover all the materials from all the components recycled, the yields are

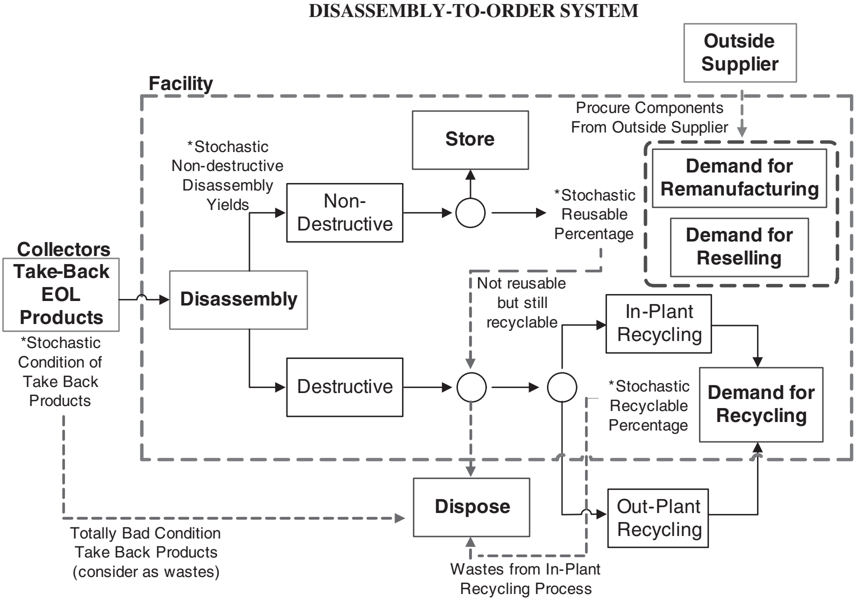

DISASSEMBLY-TO-ORDER SYSTEM
Outside Supplier
Facility
Procure Components From Outside Supplier
*Stochastic Non-destructive Disassembly Yields
Store
Demand for Remanufacturing
Demand for Reselling
Non-Destructive
*Stochastic Reusable Percentage
Collectors Take-Back EOL Products
Disassembly
Not reusable but still recyclable
In-Plant Recycling
*Stochastic Condition of Take Back Products
Destructive
*Stochastic Recyclable Percentage
Demand for Recycling
Dispose
Out-Plant Recycling
Totally Bad Condition Take Back Products (consider as wastes)
Wastes from In-Plant Recycling Process

Fig. 1. Disassembly-To-Order System.

stochastic. Components that exceed the demand and storage capacity are sent to disposal along with the waste materials from the recycling process.

HEURISTIC PROCEDURE

As mentioned before, we use one of the two heuristic approaches (viz., one-to-one approach or one-to-many approach) that convert the stochastic DTO problem into a deterministic equivalent. Here we briefly describe the two heuristics. For more details, see Inderfurth and Langella (2003).

One-to-One Heuristic

In this approach, we decompose the product into a series of single core – single component equivalents. That is, a product with multiple components is split into products with single core – single component equivalents (see Fig. 2). To do this, we split the core cost of product i (which includes the take-back cost (ctb_i) plus the separation cost (cse_i)) and estimate

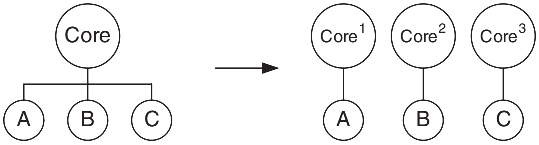


Fig. 2. Product Split into One-To-One Relation.

the core cost of component j attributed to product i ($c_{ij}{}^{S}$) in proportion to the procurement costs (cpc_j) of the components. For example, in Fig. 2, if the core cost of the product is \$9 and procurement costs of components A, B and C are \$2, \$3 and \$4 respectively, the split core costs of components A, B and C will be \$2, \$3 and \$4 respectively.

Therefore, the split core cost for component j of product i ($c_{ij}{}^{S}$) can be calculated as follows:

$$c_{ij}^{s} = (ctb_i + cse_i)\left(cpc_j / \sum_j cpc_j\right) \tag{1}$$

In order to keep the analysis simple, we assume that each stochastic yield rate (SDY_{ij}) has a continuous density function $g_{ik}(SDY_{ij})$. Therefore, each yield realization is bounded by an upper limit (SDY^{+}) and lower limit (SDY^{-}), that is, $SDY_j{}^{+} \geq SDY_j \geq SDY_j{}^{-}$.

For the uniformly distributed yield a closed-form expression for certainty equivalent for the stochastic yield rate, $S\hat{D}Y$, is given as follows (see Inderfurth & Langella, 2003):

$$S\hat{D}Y = \sqrt{\frac{cpc\,(SDY^{-})^{2} + cdc\,(SDY^{+})^{2} + 2\,c^{s}\,(SDY^{+} - SDY^{-})}{cpc + cdc}} \tag{2}$$

We use Eqs. (1) and (2) to calculate the one-to-one deterministic yield equivalent factor for each component of each product in the model.

One-to-Many Heuristic

This approach deals with the products that have single core and multiple components. Thus, there is no splitting involved in this approach. Consider a product with one core and two components as shown in Fig. 3.

We assume that the yield rates of components A and B (SDY_A and SDY_B respectively) are distributed independently. If the yield rates are uniformly

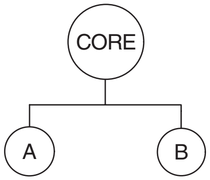


Fig. 3. Product with One Core and Two Parts.

distributed, a closed form expression for total take-back EOL products (*TTB*) can be obtained as follows (see Inderfurth & Langella, 2003).

$$TTB = \sqrt{\frac{c_{AB}\, DRC_A^2 + c_{BA}\, DRC_B^2}{c_{AB}(SDY_A^+) + c_{BA}(SDY_B^+) - 2\,(cpc_A\, S\bar{D}Y_A + cpc_B\, S\bar{D}Y_B)\,(SDY_A^+ - SDY_A^-)\,(SDY_B^+ - SDY_B^-)}} \tag{3}$$

where $c_{AB} = (cpc_A + cdc_A)\,(SDY_B^+ - SDY_B^-)$, $c_{BA} = (cpc_B + cdc_B)\,(SDY_A^+ - SDY_A^-)$, DRC_A is the demand for reuse component A, DRC_B the demand for reuse component B, $S\bar{D}Y_A$ the expected SDY_A value and $S\bar{D}Y_B$ the expected SDY_B value.

From this the certainty equivalents for the stochastic yield rates can be calculated as follows:

$$S\hat{D}Y_A = DRC_A/TTB \quad \text{and} \quad S\hat{D}Y_B = DRC_B/TTB \tag{4}$$

MATHEMATICAL MODEL

We use goal programming (GP) to determine the number of each product type to be taken back and disassembled to satisfy the demand of components and materials under a variety of physical, financial, environmental constraints and stochastic elements, while maximizing total profit and minimizing the costs of take-back EOL products, procuring the components and disposal. There are four stochastic elements in the DTO system; viz., stochastic condition of take-back products, stochastic reusable percentage, stochastic recyclable percentage and stochastic non-destructive disassembly yields. We use uniform distribution to represent the first three stochastic elements, and a heuristic procedure to represent the last stochastic element. Details of the GP model and procedure to solve the GP problem are described below.

The Goals

The first goal is to obtain the total profit of the DTO system of at least its aspiration level (TPR^*) and exceed it as much as possible. Mathematically, this can be achieved by minimizing the negative deviation (η_1) from the aspiration level (TPR^*), to maintain a value equal to zero. There is no restriction on positive deviation (ρ_1) to guarantee the best profit level. This goal can be formulated as follows:

$$\begin{aligned} \min \quad & \eta_1 \\ \text{s.t.} \quad & TPR + \eta_1 - \rho_1 = TPR^* \end{aligned} \tag{5}$$

The second goal is to limit procurement cost to have a value no more than its limitation level (CPR^*). Thus the positive deviation (ρ_2) is minimized. This goal can be formulated as follows:

$$\begin{aligned} \min \quad & \rho_2 \\ \text{s.t.} \quad & CPR + \eta_2 - \rho_2 = CPR^* \end{aligned} \tag{6}$$

The third goal is to minimize the cost of take-back products and limit it to a limitation level (CTB^*). The better combination of take-back products can reduce the total cost, while satisfying the demands for components and materials still remains. Hence, the positive deviation (ρ_3) from the take-back cost is minimized. This goal can be formulated as follows:

$$\begin{aligned} \min \quad & \rho_3 \\ \text{s.t.} \quad & CTB + \eta_3 - \rho_3 = CTB^* \end{aligned} \tag{7}$$

The forth goal is related to the environment consciousness. Our desire is to minimize the number of disposed products, components and materials by minimizing the disposal cost and limit it to a limitation level (CDI^*). Therefore, the positive deviation (ρ_4) from the disposal cost is minimized. This goal can be formulated as follows:

$$\begin{aligned} \min \quad & \rho_4 \\ \text{s.t.} \quad & CDI + \eta_4 - \rho_4 = CDI^* \end{aligned} \tag{8}$$

These goals are ordered by priority. Hence, there are no weights assigned to any of the goals. The first priority goal would be to obtain the *TPR* value of at least its aspiration level (TPR^*), while the last priority goal would be to obtain the *CDI* value with no more than its limitation level (*CDI**).

Total Profit Value and Related Terms

The total profit value can be calculated by

$$TPR = RCS + RMS + SVL - CPR - CTB - CDI \\ - CDD - CND - CST - CRE \tag{9}$$

Each term of the TPR function is described below.

RCS is calculated as a function of the total number of reused component type j (TRC_j) and the unit resale value for component type j (pc_j). Therefore,

$$RCS = \sum_j (TRC_j \, pc_j) \tag{10}$$

RMS is calculated as a function of the total weight of recycled material type j ($TWRM_j$) and the price of material type j (pm_j). Therefore,

$$RMS = \sum_j (TWRM_j \, pm_j) \tag{11}$$

SVL is calculated as a function of the total number of stored components type j (TSC_j) and the unit equivalent value for component type j (sev_j). Therefore,

$$SVL = \sum_j (TSC_j \, sev_j) \tag{12}$$

CPR is a function of the number of procurement component type j (OPC_j) and procurement cost of component type j (cpc_j). Therefore,

$$CPR = \sum_j (OPC_j \, cpc_j) \tag{13}$$

CTB is calculated as a function of the total number of EOL products type i ordered (TTB_i) and the take-back cost for each product i (ctb_i). Therefore,

$$CTB = \sum_i (TTB_i \, ctb_i) \tag{14}$$

CDI is a function of the total number of disposing product type i ($TDIP_i$), disposing component type j ($TDIC_j$) and disposing material type j ($TDIM_j$) , and their corresponding disposing cost for product type i (cdp_i), component type j (cdc_j) and material type j (cdm_j). Therefore,

$$CDI = \sum_i (TDIP_i cdp_i) + \sum_j \{(TDIC_j cdc_j) + (TDIM_j cdm_j)\} \tag{15}$$

CDD is a function of the total amount of destructive disassembly materials type j ($TDDM_j$) and the destructive disassembly cost for material type j (cdd_j). Therefore,

$$CDD = \sum_j (TDDM_j \; cdd_j) \tag{16}$$

CND is a function of the total number of non-destructive disassembly components type j ($TNDC_j$) and the non-destructive disassembly cost for component j (cnd_j). Therefore,

$$CND = \sum_j (TNDC_j \; cnd_j) \tag{17}$$

CST is a function of the total number of stored components type j (TSC_j) and the corresponding unit holding cost (cuh_j). Therefore,

$$CST = \sum_j (TSC_j \; cuh_j) \tag{18}$$

CRE is a function of the total amount of materials recycled in-plant (TIR_j) and out-plant (TOR_j), and the corresponding unit recycling cost in-plant (cir_j) and out-plant (cor_j). Therefore,

$$CRE = \sum_j \{(TIR_j \; cir_j) + (TOR_j \; cor_j)\} \tag{19}$$

Constraints

Disassembly Process

The total number of products type i to be disassembled (TDP_i) has to be equal to total number of EOL product type i (TTB_i) ordered multiplied by stochastic good condition percentage of product type i (SGP_i). Therefore,

$$TDP_i = TTB_i \; SGP_i \tag{20}$$

The total number of disassembly components and materials type j (TDC_j) has to be equal to the sum of the total number of EOL product type i (TTB_i)

multiplied by the multiplicity of component type j in product type i (CM_{ij}). Therefore,

$$TDC_j = \sum_i (TTB_i \; CM_{ij}) \tag{21}$$

The total number of non-destructive disassembly components has to be no more than the sum of the multiplication of the total number of EOL product type i (TTB_i), the multiplicity of component type j in product type i (CM_{ij}) and non-destructive disassembly yields of component type j from product type i (SDY_{ij}) calculated according to the type of heuristic approach and stored non-destructive disassembly components from previous period (PSC_j). Therefore,

$$TNDC_j \leq \sum_i (TTB_i \; CM_{ij} \; SDY_{ij}) + PSC_j \tag{22}$$

The total amount of destructive disassembly materials type j ($TDDM_j$) has to be equal to the subtraction of the total number of disassembly components type j (TDC_j) by non-destructive disassembly components type j ($TNDC_j$). Therefore,

$$TDDM_j = TDC_j - TNDC_j \tag{23}$$

Reuse Process

The number of components type j sent to reuse process (RPC_j) has to be less than or equal to the number of non-destructive disassembly of components type j ($TNDC_j$). Therefore,

$$RPC_j \leq TNDC_j \tag{24}$$

The demand of reuse component type j (DRC_j) has to be equal to the sum of good-to-reuse component type j (GRC_j) and component type j procured from outside supplier (OPC_j). Therefore,

$$DRC_j = GRC_j + OPC_j \tag{25}$$

The number of good-to-reuse components type j (GRC_j) has to be equal to the multiplication of the number of components type j sent to reuse process (RPC_j) and their stochastic reusable percentage (SRC_j). Therefore,

$$GRC_j = RPC_j \; SRC_j \tag{26}$$

The number of bad-to-reuse components type j in the reuse process (BRC_j) has to be equal to the number of components type j sent to reuse process (RPC_j) subtracted by the number of good-to-reuse component type j (GRC_j). Therefore,

$$BRC_j = RPC_j - GRC_j \tag{27}$$

The number of components type j procured from outside (OPC_j) has to be equal to the maximum of zero and the difference between the demand of reuse component type j (DRC_j) and the number of reusable component type j (GRC_j). Therefore,

$$OPC_j = \text{Max}\{0, (DRC_j - GRC_j)\} \tag{28}$$

Recycle Process

The amount of materials type j to be recycled ($TREM_j$) has to be equal to the demand of recycled materials type j ($DREM_j$). Therefore,

$$TREM_j = DREM_j \tag{29}$$

The amount of materials type j to be recycled ($TREM_j$) has to be no more than the sum of the number of bad-to-reuse components type j (BRC_j) and the total number of destructive disassembly materials type j ($TDDM_j$). Therefore,

$$TREM_j \leq BRC_j + TDDM_j \tag{30}$$

The amount of material type j to be recycled ($TREM_j$) has to be equal to the sum of the total amount of materials type j recycled in-plant (TIR_j) and out-plant (TOR_j). Therefore,

$$TREM_j = TIR_j + TOR_j \tag{31}$$

The amount of materials type j to be recycled in plant (TIR_j) has to be no more than in-plant capacity of material type j (IPC_j). Therefore,

$$TIR_j \leq IPC_j \tag{32}$$

The total weight of the recycling material type j (TRM_j) is equal to the multiplication of the number of material type j to be recycled ($TREM_j$), weight of material type j (WM_j) and its stochastic recyclable percentage (SRE_j). Therefore,

$$TWRM_j = TREM_j\ WM_j\ SRE_j \tag{33}$$

Store Process

The total number of stored component type j (TSC_j) is equal to the difference between the number of non-destructive disassembly component type j (NDC_j) and the number of components type j sent to reuse process (RPC_j). Therefore,

$$TSC_j = NDC_j - RPC_j \tag{34}$$

The total number of stored component type j (TSC_j) has to be no more than available storage capacity for component type j (ASC_j). Therefore,

$$TSC_j \leq ASC_j \tag{35}$$

Disposing Process

The total number of disposing product type i ($TDIP_i$) is equal to the difference between the total number of EOL product type i ordered (TTB_i) and the number of product type i to be disassembled (TDP_i). Therefore,

$$TDIP_i = TTB_i - TDP_i \tag{36}$$

The total number of disposing component type j ($TDIC_j$) is equal to the sum of the number of bad-to-reuse components type j in the reuse process (BRC_j) and the total amount of destructive disassembly materials type j ($TDDM_j$), subtracted by the demand for recycle materials type j ($DREM_j$). Therefore,

$$TDIC_j = BRC_j + TDDM_j - DREM_j \tag{37}$$

The total amount of disposing material type j ($TDIM_j$) is equal to the multiplication of the total amount of material type j recycled in-plant (TIR_j), weight of material type j (WM_j) and non-recyclable percentage ($1-SRE_j$). Therefore,

$$TDIM_j = TIR_j\ WM_j(1 - SRE_j) \tag{38}$$

$$\begin{aligned}&\text{Variables } TTB_i,\ \ TDIP_i,\ \ TNDC_j,\ \ RPC_j \text{ and } GRC_j \text{ must be}\\&\text{integer, and all variables must be non-negative.}\end{aligned} \tag{39}$$

The GP Model

The GP model can now be written as follows:

Find (TTB_j, $TNDC_j$, $TDDM_j$, $\underline{TRC_j}$, $TREM_j$, TSC_j, and $TDIC_j$,) so as to:

$$\text{Lexicographically minimize } u = \{(\eta_1),\ (\rho_2),\ (\rho_3),\ (\rho_4)\} \tag{40}$$

where $\{\eta_k, \rho_k\}$ are defined in Eq. (5) through Eq. (8) (the terms of which are explained in Eq. (9) through Eq. (19)).

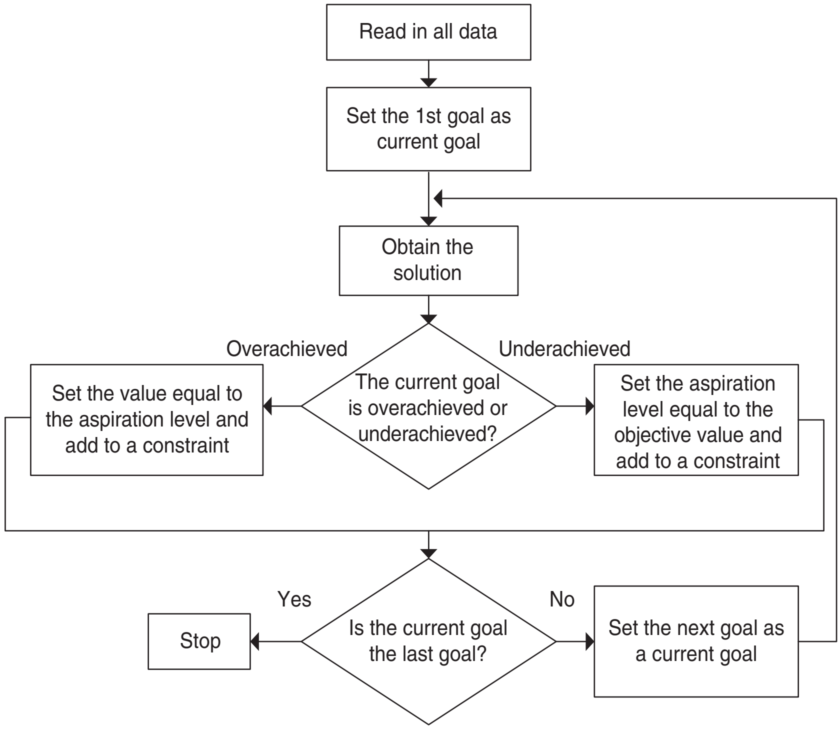

Read in all data
Set the 1st goal as current goal
Obtain the solution
Overachieved
Underachieved
The current goal is overachieved or underachieved?
Set the value equal to the aspiration level and add to a constraint
Set the aspiration level equal to the objective value and add to a constraint
Yes
No
Is the current goal the last goal?
Stop
Set the next goal as a current goal

Fig. 4. Procedure to Solve GP Model.

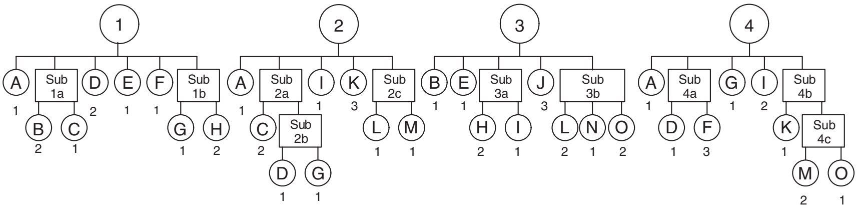


Fig. 5. Product Structures for Case Example.

Subject to:
Constraints defined in Eq. (20) through Eq. (39) and $\{\eta_k, \rho_k\} \geq 0, k = 1, \ldots, 4.$

Procedure to Solve the GP Model

The procedure to solve the GP model is as shown in Fig. 4.

Table 1. Input data for the Case Example.

Component$_j$	Price of Component	Price of Material	Demand for Reuse Comp.	Demand for Recycle Mat.	Unit Holding Cost	Store Value	Dest. Diss. Cost	Non-Dest. Diss. Cost	Disposing Cost	Material Disposing Cost	Outside Procurement Cost	Stochastic Reusable Percentage	Stochastic Recyclable Percentage	Weight of Comp.
A	12	7	400	250	0.45	10	0.35	0.55	0.6	0.4	12	0.8	0.8	1.2
B	11	5	400	250	0.35	10	0.35	0.55	0.4	0.2	11	0.85	0.9	1.6
C	10	7	350	200	0.5	9	0.25	0.5	0.7	0.5	10	0.85	0.7	1.4
D	9	8	450	250	0.45	8	0.3	0.6	0.7	0.5	9	0.75	0.85	1.2
E	12	8	350	220	0.55	11	0.35	0.55	0.8	0.6	12	0.8	0.9	1.3
F	10	6	300	250	0.65	9	0.3	0.4	0.4	0.2	10	0.8	0.7	0.8
G	11	6	450	220	0.55	9	0.25	0.6	0.9	0.7	11	0.9	0.65	1.1
H	10	4	400	220	0.55	9	0.4	0.5	0.7	0.5	10	0.95	0.85	1.5
I	13	6	450	280	0.4	12	0.3	0.55	0.6	0.4	13	0.8	0.8	1.2
J	10	6	300	200	0.4	9	0.45	0.5	0.8	0.6	10	0.85	0.75	0.8
K	11	7	350	220	0.35	10	0.25	0.4	0.9	0.7	11	0.9	0.7	1.6
L	11	8	400	250	0.5	10	0.4	0.55	0.5	0.7	11	0.95	0.85	0.9
M	12	8	400	200	0.5	10	0.3	0.45	0.8	0.6	12	0.75	0.85	0.9
N	11	6	300	220	0.45	10	0.3	0.4	0.8	0.6	11	0.8	0.9	0.9
O	12	7	350	250	0.5	11	0.35	0.5	0.9	0.7	12	0.7	0.75	1.1

Note: We use $U(0.9,1)$ for SGP_i value, $U(0.7,1)$ for SRC_j value and $U(0.65,0.9)$ for SRE_j value.

CASE EXAMPLE

In this example, we have 4 products and each of them has 8 components. The structures of products are shown in Fig. 5. Also, the component multiplicity for each component is shown below each component in Fig. 5. The input data are given in Table 1.

Additional data needed for the case example are: product take-back cost $(ctb_i) = 35, 32, 35, 34$, product disposal cost $(cdp_i) = 4, 4, 4, 4$, stochastic percentage of good product $(SGP_i) = 0.91, 0.95, 0.92, 0.96$ for product i, $i = 1, 2, 3$ and 4, in-plant recycling cost $(cir_j) = 1$, out-plant recycling cost $(cor_j) = 1.8$, in-plant capacity $(IPC_j) = 120$, available storage space $(ASC_j) = 20$ for each of the component type j, $SDY_A = SDY_B = 0.55$, $SDY^+ = 0.9$, $SDY^- = 0.3$, aspiration levels for total profit, procurement cost, take-back cost and disposing cost are 27,000, 1,500, 15,000 and 1,000, respectively.

From Eqs. (1) through (4), we can calculate the stochastic non-destructive disassembly yield value of each component for each product as shown in Table 2.

We used LINGO 7 to calculate the maximum profit with one-to-one and one-to-two approaches. The maximum profit obtained using the one-to-one approach was \$27,780.20, while using the one-to-many approach was \$28,094.70.

One-to-many approach generates more profit than one-to-one approach. Therefore we select one-to-many approach. When applied to the goal-programming model, we obtain the results shown in Table 3.

Table 2. Deterministic Yield Equivalents for One-to-One and One-to-Many Approaches.

Product 1			Product 2			Product 3			Product 4		
Items	One-to-One	One-to-Many	Items	One-to-One	One-to-Many	Items	One-to-One	One-to-Many	Items	One-to-One	One-to-Many
A	0.737	0.436	A	0.725	0.454	B	0.752	0.799	A	0.727	0.801
B	0.727	0.959	C	0.723	0.953	E	0.766	0.862	D	0.722	0.760
C	0.736	0.547	D	0.720	0.737	H	0.755	0.959	F	0.722	0.981
D	0.732	0.943	G	0.725	0.808	I	0.756	0.598	G	0.727	0.368
E	0.742	0.876	I	0.720	0.346	J	0.761	0.865	I	0.722	0.928
F	0.732	0.682	K	0.731	0.938	L	0.751	0.789	K	0.733	0.419
G	0.737	0.527	L	0.713	0.748	N	0.769	0.792	M	0.729	0.963
H	0.731	0.957	M	0.726	0.798	O	0.766	0.865	O	0.733	0.405

Table 3. Values of Various Variables at Different Steps of the Procedure to Solve the GP Model.

Goals	Aspiration Level	1st Run	Goal 1	Goal 2	Goal 3	Goal 4
Total profit variable	≥27,000	28,094.7	27,000	27,000	27,000	27,000
Procurement cost	≤1,500	2,491	1,289	1,493	1,493	1,493
Take-back EOL products cost	≤15,000	13,190	14,830	14,370	14,420	14,420
Disposing cost	≤1,000	641.2	1,132.648	1,050.265	940.91	940.91

The numbers of optimal take-back EOL products to satisfy all goals for products 1, 2, 3 and 4 are 220, 160, 180 and 200 respectively.

CONCLUSIONS

In this paper, we implemented two heuristic approaches to handle the stochastic elements of the DTO system. Using and comparing the results obtained from one-to-one and one-to-many heuristic approaches helped us to choose the effective deterministic yield equivalents. We used a GP procedure to determine the number of returned products that satisfy the four goals. When solved, all aspirations were satisfied without violating the constraints.

REFERENCES

Gungor, A., & Gupta, S. M. (1999). Issues in environmentally conscious manufacturing and product recovery: A survey. *Computers and Industrial Engineering*, *36*(4), 811–853.

Imtanavanich, P., & Gupta, S. M. (2004). Multi-criteria decision making for disassembly-to-order system under stochastic yields, *Proceedings of the SPIE International Conference on Environmentally Conscious Manufacturing IV*, Philadelphia, Pennsylvania, October 26–27, (pp. 147–162).

Inderfurth, K., & Langella, I. M. (2003). An approach for solving disassemble-to-order problems under stochastic yields. *Proceeding of Logistics Management*, Braunschweig, Germany, September 24–26.

Kongar, E., & Gupta, S. M. (2002). A multi-criteria decision making approach for disassembly-to-order system. *Journal of Electronics Manufacturing*, *11*(2), 171–183.

Lambert, A. J. D., & Gupta, S. M. (2002). Demand driven disassembly optimization for electronic products. *Journal of Electronics Manufacturing*, *11*(2), 121–135.

Lambert, A. J. D., & Gupta, S. M. (2005). *Disassembly modeling for assembly, maintenance, reuse, and recycling*. Boca Raton, FL: CRC Press ISBN: 1-57444-334-8.

Moyer, L., & Gupta, S. M. (1997). Environmental concerns and recycling/disassembly efforts in the electronics industry. *Journal of Electronics Manufacturing*, *7*(1), 1–22.

Sodhi, M. S., & Reimer, B. (2001). Models for recycling electronics end-of-life products. *OR Spektrum*, *23*(1), 97–115.

Stuart, J. A., & Qin, L. (2000). A model for discrete processing decisions for bulk recycling of electronics equipment. *IEEE Transactions on Electronics Packaging Manufacturing*, *23*(4), 314–320.

Veerakamolmal, P., & Gupta, S. M. (1998). Optimal analysis of lot size balancing for multi-products selective disassembly. *International Journal of Flexible Automation and Integrated Manufacturing*, *6*(3/4), 245–269.

Veerakamolmal, P., & Gupta, S. M. (1999). Analysis of design efficiency for the disassembly of modular electronic products. *Journal of Electronics Manufacturing*, *9*(1), 79–95.

This page intentionally left blank

PART III: OPERATIONAL APPLICATIONS

This page intentionally left blank

SELECTION CRITERIA FOR A NETWORK DESIGN MODEL WITH UNCERTAIN DEMAND

Steven Cosares and Fred J. Rispoli

ABSTRACT

We address the problem of selecting a topological design for a network having a single traffic source and uncertain demand at the remaining nodes. Solving the associated fixed charge network flow (FCF) problem requires finding a network design that limits both the fixed costs of establishing links and the variable costs of sending flow to the destinations. In this paper, we discuss how to obtain a sequence of optimal solutions that arise as the demand intensity varies from low levels to high. One of the network design alternatives associated with these solutions will be chosen based upon the dominant selection criteria of the decision maker. We consider both probabilistic and non-probabilistic criteria and compare the network designs associated with each. We show that the entire sequence of optimal solutions can be identified with little more effort than solving a single FCF problem instance. We also provide solution approaches that are relatively efficient and suggest good design alternatives based upon approximations to the optimal sequence.

Applications of Management Science: In Productivity, Finance, and Operations
Applications of Management Science, Volume 12, 129–143

ISSN: 0276-8976/doi:10.1016/S0276-8976(06)12010-6

1. INTRODUCTION AND NOTATION

Owing to their importance in telecommunication, transportation, power delivery, etc., the problem of network design has become increasingly important in our highly connected world. General network design models (see, e.g., Ahuja, Magnanti, & Orlin, 1993; Annals of Operations Research, 2001; Anandalingam & Raghavan, 2002; Balakrishnan, Magnanti, Shulman, & Wong, 1991; Barnhart, Krishnan, Kim, & Ware, 2002; Magnanti, Wolsey, & Wong, 1992; Sherali, Subramanian, & Loganathan, 2001), consist of finding some subset of a given network to support the demand for flow between pairs of locations. Each potential link in the network might have associated with it a "fixed" cost, which is incurred if the link is utilized, a "variable" cost that depends on the number of units of flow it supports, and a capacity on its total flow. A design objective would be to find a subnetwork that can feasibly support the demands at minimum total cost. Computationally, these are often hard problems, requiring some form of enumeration in order to guarantee optimality in the solution. In some cases, the decision maker might be willing to settle for some less-than-optimal design that could be more easily obtained if it is believed to have a relatively low cost and would satisfy additional design criteria regarding, e.g., network reliability, demand protection, and insensitivity to fluctuations in level of demand.

General models for network design might have multiple source and destination locations for the commodity being shipped, or might involve multiple commodities to be delivered between pairs of locations over a shared infrastructure. A fairly exhaustive list of the basic heuristic algorithms to find reasonable designs is presented in Minoux (1989). Most studies regarding building a network to satisfy uncertain demands (see, e.g., Infager, 1994; Riis & Andersen, 2002; Sen, Doverspike, & Cosares, 1994), focus on establishing appropriate capacity levels for an existing network, rather than on topological design, which is the focus of this paper.

Depending on the intensity of the demand, the network designer must trade off the fixed costs of building the network with the variable costs of shipping flow over its links. At one extreme, when the level of demand for service is very high, the emphasis might be placed on finding a design that minimizes the (variable) transmission costs only. This would be appropriate if the fixed cost of the links were small in comparison. At the other extreme, when the demand is low, the emphasis would be on minimizing the total fixed cost of the network. The resulting design, however, might be very expensive if the demand turns out to be higher than expected as the variable link costs would then become more dominant. For intermediate levels of demand, the (elusive) network design is the one that represents the optimal balance of fixed and variable costs. The task of finding such a design is classified as "NP-Complete". It is even more daunting when the design decisions must be made before the actual level of demand intensity is known which is often true.

In this paper, we consider networks having a single source location for the commodities, node 1 and uncertain demand at the remaining nodes $2, \ldots, n$. We assume that the links can support an unlimited amount of flow. Such cases still exhibit the computational complexity of the more general network design model, even though the optimal design is known to take the form of some spanning tree of the network (see Hochbaum & Segev, 1989). In this case, very low levels of demand would be best served with a tree that minimizes the total fixed costs – a minimum spanning tree (MST), while very large levels of demand would require a design made up of minimum variable cost paths from node 1 to every other node – a shortest paths tree (SPT). The optimal designs at these extremes are computationally easy to obtain (see, e.g., Horowitz, Sahni, & Rajasekaran, 1997; Tarjan, 1983).

1.1. Mathematical Programming Formulation

When the demand is known, the optimal network design can be found by solving an instance of the *fixed charge network flow* (FCF) *problem* over an undirected network with n nodes. Let f_{ij} represent the non-negative fixed cost of activating the link between nodes i and j, and let $c_{ij} = c_{ji}$ represent the non-negative cost per unit flow (in either direction) over link $\{i,j\}$. We let y_{ij} be a binary decision variable that is set to 1 if link $\{i,j\}$ is included in the network design, and $y_{ij} = 0$ otherwise. Let $(x_{ij}+x_{ji})$ represent the total amount of flow (in either direction) that traverses link $\{i,j\}$, if it is included in the optimal design, to satisfy the positive demands d_j at destination nodes $j = 2, \ldots, n$.

We introduce an "intensity parameter" β to denote the overall magnitude of network demand. We assume that its precise value is not known at the time the network is being designed. Each of the demands d_j is scaled by this factor to represent the case where the relative demands across the destination nodes is given, but their relative level of intensity is not. This would model the situation where some new commodity/service is planned for distribution to locations in the network having predictable relative demand patterns (e.g., based on history), but where the overall popularity of this new commodity/service cannot be precisely forecast. Some probability distribution data regarding the value of β might be available. The FCF could be represented as a mixed integer linear program as follows.

Minimize:

$$\sum_{i=1}^{n}\sum_{j=i+1}^{n} f_{ij}y_{ij} + \sum_{i=1}^{n}\sum_{j=2}^{n} c_{ij}x_{ij}$$

Subject to:

$$\sum_{p} x_{pj} - \sum_{q} x_{jq} = d_j * \beta \quad j = 2, \ldots, n \tag{1}$$

$$My_{ij} \geq x_{ij} + x_{ji}, \qquad i = 1, \ldots, n; \quad j = i, \ldots, n \tag{2}$$

$$y_{ij} \in \{0, 1\}, \qquad i = 1, \ldots, n; \quad j = i, \ldots, n \tag{3}$$

$$x_{ij} \geq 0, \qquad i = 1, \ldots, n; \quad j = 2, \ldots, n \tag{4}$$

Constraints (1) represent flow balance constraints, requiring the residual flow at each node to be equal to the (scaled) demand. Constraints (2) ensure that if y_{ij} is 0 then no flow will pass link $\{i,j\}$ in either direction. The value of M is set large enough to guarantee that flow levels are unconstrained for included links. For any positive fixed value of β, the optimal y variables that have value 1 define the optimal (tree) network design.

1.2. Tree-Based Cost Functions

For a spanning tree T, let $F(T)$ be the sum of all f_{ij}, where $\{i,j\} \in T$, i.e., the total fixed cost of the design. Let $C(T,j)$ be the product of d_j and the sum of the c_{ij} in the unique path in T from source node 1 to destination node j. Let $V(T)$ be the sum of the $C(T,j)$, i.e., the total variable (flow) cost associated with tree T. The objective in FCF with demand intensity parameter β is to find a spanning tree T that minimizes the cost function $\phi(T, \beta) = F(T) + \beta * V(T)$.

Any spanning tree that is an optimal design for some positive range of values of β is called an *optimal tree*. We define MST to be a MST that is based solely on fixed costs f_{ij}. In cases where this tree is not unique, we select one that has the smallest value of $V(T)$. MST is an optimal tree when the demand intensity is expected to be very low. We define SPT to be a spanning tree consisting of the shortest paths from source node 1 to every other node, based solely on the variable costs c_{ij}. In cases where there is more than one such tree, we select one that has the smallest value of $F(T)$. SPT is an optimal tree when the demand intensity is expected to be sufficiently high. A number of network design approaches (see, e.g., Khuller, Raghavachari, & Young, 1995) involve building solutions for intermediate demand levels that represent some compromise between these two extreme designs. When the demand is uncertain, it may be necessary to examine the behavior of FCF as the intensity parameter β varies, i.e., to look at the continuum of optimal tree designs between MST and SPT. Using decision analysis, a specific design decision could be made by selecting one of these optimal tree solutions.

We define $\phi*(\beta)$ to be the total cost of an optimal tree associated with demand intensity $\beta\cdot$ So if T is an optimal tree, then $\phi*(\beta) = \phi(T, \beta)$ for some range of $\beta\cdot$ The function $\phi*(\beta)$ together with the sequence of optimal trees that define $\phi*$ comprise what we call the *optimal spectrum* for the instance of FCF. Consider, for example, the FCF given in Fig. 1.

Suppose, without loss of generality, that $d_j = 1$ for $j = 2,3,4$. The above network contains eight spanning trees; each with a cost function $\phi(T,\beta)$. In this case, the functions are: $17 + 12\beta$ (for MST with links {1,2}, {2,4}, {3,4}), $18 + 8\beta$, $20 + 7\beta$, $21 + 6\beta$, $21 + 12\beta$, $25 + 8\beta$, $24 + 6\beta$, and $25 + 5\beta$ (for SPT with links {1,2}, {1,3}, {1,4}). Only four of these eight functions is minimum for some range of values of β; these correspond to the sequence of optimal trees, which begins with MST and ends with SPT. The remaining four trees can be removed from consideration. The optimal objective function $\phi*(\beta)$ for this instance can be illustrated with the graph given in Fig. 2.

The function $\phi*(\beta)$ is defined by the lower boundary of the set of curves. We define a *breakpoint* to be a value of β where two different optimal trees define $\phi*(\beta)$. The spectrum for the network in Fig. 1 can be represented as

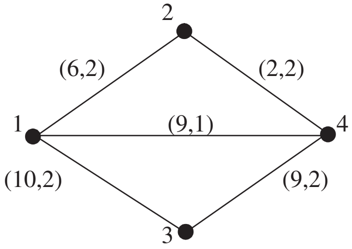


Fig. 1. A Simple Instance of FCF. The Link Costs are (f_{ij}, c_{ij}).

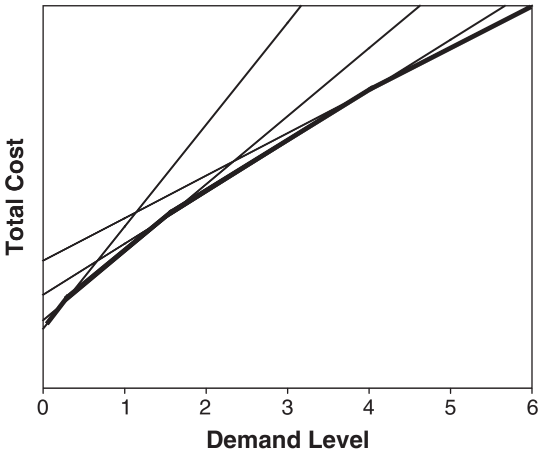


Fig. 2. The Graph of $\phi*(\beta)$ is in Bold.

follows:

$$\phi*(\beta) = \begin{cases} 17 + 12\beta, & \text{for } 0 \leq \beta \leq .25, \quad \text{where } T_1 = \text{MST} = \{\{1,2\},\{3,4\},\{2,4\}\} \\ 18 + 8\beta, & \text{for } 0.25 \leq \beta \leq 1.5, \quad \text{where } T_2 = \{\{1,2\},\{1,3\},\{2,4\}\} \\ 21 + 6\beta, & \text{for } 1.5 \leq \beta \leq 4.0, \quad \text{where } T_3 = \{\{1,3\},\{1,4\},\{2,4\}\} \\ 25 + 5\beta, & \text{for } 4.0 \leq \beta, \quad \text{where } T_4 = \text{SPT} = \{\{1,2\},\{1,3\},\{1,4\}\} \end{cases}$$

Based upon the available information regarding potential values for β and the prominent selection criteria of the decision maker, one of these optimal trees would be selected for the ultimate network design.

The remainder of the paper is organized as follows. In Section 2, we describe properties of $\phi*$ as a function of β. We present an $O(k \log k)$ algorithm that identifies the optimal spectrum, where k is the number of spanning trees in the network. We show that the work required to find the entire sequence of solutions is no more than a factor of $O(\log n)$ over any enumerative method that solves FCF for a single specific demand. In Section 3, we discuss how various decision-making criteria could be applied to select a reasonable design from the set of optimal trees. In Section 4, we describe some schemes that efficiently generate approximations to the optimal spectrum. The spanning trees obtained from these schemes could be used as decision alternatives from which to select a network design. A comprehensive example is provided to illustrate the application of these schemes.

2. CHARACTERISTICS OF THE OPTIMAL SPECTRUM

We say that a non-empty subset of trees $\{T_j\}$ *dominates* tree T if for every feasible value of β there is a T_j from the subset such that $\phi(T_j, \beta) < \phi(T, \beta)$. In this case T can never be an optimal tree. This occurs in the above example for the tree T with $\phi(T, \beta) = 20 + 7\beta$; it is dominated by $\{T_1, T_2, T_3\}$. A spanning tree T is called *dominant* if it, by itself, dominates every other tree. In this case it would define the function $\phi*$ for the entire range of β. For example, if all of the f_{ij} are equal, then SPT would be dominant. If c_{ij} are all zero, then MST would be dominant.

Suppose, more generally, that neither MST nor SPT is dominant. Then it would follow that the spectrum has at least one breakpoint. The cost function associated with any of the remaining spanning trees in the network has the potential for defining some segment of $\phi*(\beta)$. For a network with n nodes and k spanning trees, algorithms are known that find all spanning trees in $O(kn)$ time (see, e.g., Gabow & Myers, 1978; Matsui, 1997). Any such algorithm could be used in our proposed process to construct the optimal spectrum.

Theorem 1. For any instance of FCF over a network having k spanning trees.

(a) The optimal cost function, $\phi*(\beta)$, is concave piecewise linear function whose segments are defined by a sequence of optimal trees $T_1 \ldots T_N$, where T_1 is MST and T_N is SPT·

(b) The sequence of optimal trees satisfies $F(T_1) < F(T_2) < \cdots < F(T_N)$ and $V(T_1) > V(T_2) > \cdots > V(T_N)$.

Proof. (a) Consider the linear functions $\phi(T, \beta) = F(T) + \beta V(T)$, where T is a spanning tree. Since $\phi*(\beta)$ is defined to be the minimum of these functions, it must be a piecewise linear concave function. The optimal trees that correspond to the minimums for some range of β can be ordered and numbered $T_1, T_2, \ldots, T_N$, according to where in the range their functions define $\phi*$. Observe that when β is very small, T_1 corresponds to the function with the smallest value of $F(T)$, i.e., MST. For large enough β, T_N would correspond to the network with the smallest $V(T)$, i.e., SPT. (b) Let T_i and T_{i+1} be two distinct optimal trees. Because $\phi*$ is concave, it follows that the slopes $V(T_i) > V(T_{i+1})$. Because both trees are optimal, there must exist a breakpoint value $\beta > 0$ where $\phi(T_i, \beta) = F(T_i) + \beta V$

$(T_i) = \phi(T_{i+1},\beta) = F(T_{i+1}) + \beta V(T_{i+1})$. Since $\beta V(T_i) > \beta V(T_{i+1})$, it must follow that $F(T_i) < F(T_{i+1})$.

Theorem 2. The entire spectrum of the FCF can be obtained in $O(k \log k) = O(kn \log \mathrm{n})$ time.

The proof is provided in the appendix. Cayley's theorem states that the number of spanning trees in a network is bounded by n^{n-2}, so the term $\log k = O(\mathrm{n} \log \mathrm{n})$. Thus the method to find the spectrum has complexity that is no more than a factor of $\mathrm{O}(\log n)$ over some $O(kn)$ tree enumeration scheme that solves an instance of FCF with fixed β.

Clearly the number of optimal trees N in $\phi*(\beta)$ is bounded by the number of spanning trees k in the network. Observe that N can serve as a measure indicating how often the optimal network design changes as the demand intensity varies. One could use this measure to gauge the relative sensitivity of the parametric FCF instance to changes in the demand intensity. In general, the number of optimal trees, hence N, is limited by $O(n^{n-2})$. However, for some instances N could be considerably smaller, i.e., only some proper subset of spanning trees ever defines the optimal spectrum. We leave as a challenging open question whether instances over a complete network exist in which every spanning tree is optimal for some range of $\beta\cdot$ If no such instance exists, then could a tighter general bound on N be derived? One class of instances with a tight bound on N are those in which the set of distinct values for $F(T)$ (or $V(T)$) lie within some narrow range. In this case, the number of distinct functions defining the spectrum may be very small.

Theorem 3. Let $\Delta F = F(\mathrm{SPT}) - F(\mathrm{MST}) + 1$ and $\Delta V = V(\mathrm{MST}) - V(\mathrm{SPT}) + 1$. An instance of FCF with integer f_{ij} and c_{ij}, has no more than $\min(\Delta F, \Delta V)$ optimal trees.

Proof. Suppose $\Delta V \leq \Delta F$ (the argument is similar when $\Delta F \leq \Delta V$). Consider the set of all spanning trees $\{T_1, T_2, \ldots, T_k\}$, and define the relation $T_i \sim T_j$ if and only if $V(T_i) = V(T_j)$. Clearly $\sim$ is an equivalence relation. Since the c_{ij} are integers, the number of equivalence classes is at most ΔV. Moreover, for each class there is a spanning tree whose total fixed costs is minimal amongst all other trees in its class. Therefore, the spectrum can be defined using at most ΔV distinct linear functions. This implies at most ΔV optimal trees.

Corollary. If $c_{ij} = c$ for all $\{i,j\}$ then the number of optimal trees is $O(n^2)$.

The result holds in the above case because the number of distinct values for $V()$ is bounded by $(n)(n\text{-}1)/2$.

3. SELECTING A NETWORK DESIGN

In addition to providing insight about the relative sensitivity of a problem instance to the demand intensity, the optimal spectrum indicates which spanning trees are optimal for some β and hence are good candidates for the network design to be selected and ultimately built. This decision would be based on the available information regarding the likely values for β and the selection criteria applied. For example, if a valid probability distribution for β were available then a decision maker seeking the design having the minimum expected total cost would select the optimal tree associated with the expected value of β. Another decision maker might select the tree that is optimal for the largest range of feasible β or for the range having the largest total probability.

If some discrete set of possible values $\beta_1, \ldots, \beta_m$ were established, e.g., a minimum, a maximum, and some potential intermediate values, then a decision table model could be built. Let $T_1, \ldots, T_m$ be the set of optimal trees (decision alternatives) associated with these values for β. If some T_i occurs for more than one of the β_j then the redundant information may be removed. The *ij*th entry in the cost matrix, denoted by Φ, is obtained by calculating the value of $\phi(T_i,\beta_j)$. Then we can select a design based on some decision criteria. For example, the *minimax* design is the tree that minimizes the maximum outcome for each tree T_i. One can apply the *equally likely* or *Laplace* criterion which finds the design with the smallest average cost. In fact any of the decision-making criteria under uncertainty such as those discussed in standard quantitative management texts can be used (e.g., see Render & Stair, 1991).

If, in addition, probability values for $\beta_1, \ldots, \beta_m$ are known, then using Φ one could determine a *minimum expected cost* design. Alternatively, some risk-based criterion could be applied. It is evident from observing the spectrum that the designs that pose the minimum financial risk, i.e., that are least sensitive to fluctuations in β, would be those whose cost functions have the smallest slopes, $V(T)$. Hence, decision makers who are risk averse are more likely to select the SPT design, even if it is likely to be more expensive than an intermediate optimal tree from the spectrum. In fact, the *minimax* design is always the SPT if it is among the alternative designs. On the other hand, the spectrum indicates that the MST, which is the design of choice if demands are very small, carries the greatest financial risk, since its cost function has the greatest slope among the optimal trees. A comprehensive example illustrating the above criteria is given at the end of Section 4.

4. EFFICIENT SCHEMES FOR GENERATING DESIGN ALTERNATIVES

In many situations it may be impractical to build the entire optimal spectrum using the Tree-based methodology described in Section 2. For example, if one were to select the design based on the *minimum expected cost* criterion, then he or she would only need to solve the single FCF instance associated with the expected value of β, without considering any other portion of the spectrum.

If the decision maker wishes to apply a criterion that requires the entire optimal spectrum, but is convinced that the number of optimal trees, N, were relatively small (e.g., less than $O(\log n)$), then the "Breakpoint" method could be used. It makes a limited number of calls to the subroutine, $ND(B)$, which finds the optimal solution to a single instance of FCF with β set to some fixed value B. Such a subroutine might be available from a vendor of special purpose mathematical programming software.

> *The Breakpoint method.* Consider the approximate spectrum comprised of the cost functions $\phi(T, \beta)$ associated with the (easily obtained) optimal trees, MST and SPT. If neither tree is dominant, then identify the "breakpoint" value $\beta = B_1$ for which the two cost functions intersect. Run $ND(B_1)$ to identify the optimal tree T_1 associated with this breakpoint. If the tree T_1 is new to the approximate spectrum, then its cost function can be included because it is minimum for some range of β. Adding this new cost function might generate as many as two new breakpoints with the present approximate spectrum. The routine $ND()$ could now be run for these breakpoint values, potentially generating new optimal trees, new cost functions, and new breakpoints. This process would continue until no new optimal trees (hence no new breakpoints) are identified, which would indicate that the optimal spectrum has been found.

Since each new optimal tree is associated with no more than two runs of $ND()$, the overall complexity of this method is bounded by $O(kn^*N)$. One might chose to apply this method for more general problem instances with potentially high N, in which case a very large number of calls to $ND()$ might be made before identifying the entire optimal spectrum. Hence, in general, this method is much less efficient than the Tree-based method. Notice though, that if the method were terminated early, e.g., after a fixed number of calls to $ND()$, before all of the breakpoints generated were examined, then the set of optimal cost functions identified comprises an (upper-bound) approximation to the actual optimal spectrum. The decision maker could use the (sub)set of optimal trees thus far identified as the alternatives from which to select a network design.

An alternative approach is to run *ND*() for a fixed number of equally spaced values in the feasible range of β. The cost functions from the optimal trees identified provides a different approximation to the optimal spectrum, which might cover regions of values overlooked by the Breakpoint Method. We point out that once an SPT arises as the optimal tree for some value of β, then it would not be necessary to solve for any larger values of β. If there is probability distribution data for β available, then one could generate a greater number of calls to *ND*() for values of β having a higher probability of occurring, e.g., by using Monte Carlo simulation. Again, the approximation to the spectrum developed from these solutions would look much like the actual spectrum, except that some of the defining cost functions might be missing. Depending on the number of runs performed, the design selected by this approach has a reasonable likelihood of being the actual optimal or one of its close runners-up.

A drawback to the Tree-based method described in Section 2 is that it may be computationally prohibitive to identify all of the spanning trees in order to build the optimal spectrum from which to select a network design. A similar drawback to the Breakpoint method and its variants is that they rely on calls to *ND*(), which might also involve some form of enumeration. We point out that if some small subset of the spanning trees is considered, then the Tree-based method could be used to identify a sequence of "best" trees in $O(m \log m)$ time, where m is the number of trees considered. The minimum of the cost functions associated with these best trees would provide an approximation to the actual spectrum and could be used as the set of alternatives from which to select a network design. For example, the decision maker could limit attention to those spanning trees having no more than some fixed depth, e.g., 3 or 4. These can be found in polynomial time. This particular subset is quite likely to contain many of the trees from the actual optimal spectrum, especially for larger values of β, where shallower trees are more likely to be optimal. Thus, the decision maker is likely to select the same design using this method as he or she would by working with the actual optimal spectrum. Vasko et al. (2002), suggest an alternative approach to generating trees (and the associated approximation to the optimal spectrum) that essentially begins with MST and then performs "pivots" that replace one tree edge with another, possibly ending with SPT.

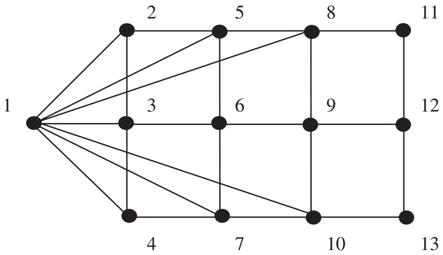


Fig 3. A Comprehensive Example.

Table 1. Fixed and Variable Costs for the Links in Fig. 3.

Link $\{i,j\}$	Cost (f, c)	Link $\{i,j\}$	Cost (f, c)	Link $\{i,j\}$	Cost (f, c)	Link $\{i,j\}$	Cost (f, c)
{1,2}	(78, 15)	{1,10}	(385, 87)	{5,6}	(58, 75)	{8,11}	(138, 38)
{1,3}	(234, 18)	{2,3}	(18, 13)	{5,8}	(73, 43)	{9,10}	(99, 47)
{1,4}	(196, 29)	{2,5}	(48, 24)	{6,7}	(204, 70)	{9,12}	(125, 65)
{1,5}	(16, 22)	{3,4}	(170, 85)	{6,9}	(195, 85)	{10,13}	(156, 65)
{1,7}	(212, 28)	{3,6}	(270, 12)	{7,10}	(93, 53)	{11,12}	(248, 15)
{1,8}	(356, 56)	{4,7}	(97, 46)	{8,9}	(137, 96)	{12,13}	(131, 23)

To illustrate some of the above approaches, consider the network given in Fig. 3. There are 13 nodes; some spanning tree, T^*, consisting of 12 of the 24 eligible links must be selected. The fixed and variable costs associated with the instance are given Table 1.

Suppose we generate values for β in equal intervals, e.g., $\beta = 2, 4, 6, 8, 10$. By running *ND*() for each value, four distinct optimal trees are identified. We know that this is a subset of the optimal trees because MST is not among this set. As a matter of fact, had we run *ND*() for $\beta = 0,1,2, \ldots, 10$, we would have identified seven optimal trees; using smaller intervals might have uncovered even more. From the set of four, we obtain the following approximate spectrum:

$$\phi(\beta) = \left\{ \begin{array}{ll} 1261 + 900\beta, & \text{for } 0 < \beta \leq 3.05 : T_1 \\ 1572 + 798\beta, & \text{for } 3.05 \leq \beta \leq 7.86 : T_2 \\ 1855 + 762\beta, & \text{for } 7.86 \leq \beta \leq 9.45 : T_3 \\ 25 + 5\beta, & \text{for } 9.45 \leq \beta : T_4 = \text{SPT} \end{array} \right\}$$

Notice that it would not be necessary to run *ND*() for values greater than 10 because SPT must be optimal for all β greater than 9.45. In this case, the

Table 2. The Cost Matrix Φ.

	Demand Intensity, β				
Alternatives	2	4	6	8	10
T_1:$1{,}261+900\beta$	3,061	4,861	6,661	8,461	10,261
T_2:$1{,}572+798\beta$	3,168	4,764	6,360	7,956	9,552
T_3:$1{,}855+762>$	3,379	4,903	6,427	7,951	9,475
T_4:$2{,}167+729>$	3,625	5,083	6,541	7,999	9,457

tree that is optimal for the largest subrange is $T_{2.}$ This would also be the tree selected if one were to apply the *Laplace* decision criterion, which assumes all values of β in the range are equally likely. By calculating the values of $\phi(T_i,\beta_j)$ for the equally spaced values, we obtain the cost matrix in Table 2.

The *minimax regret* criterion would suggest design T_2. Observe that, as expected, the *minimax* design is T_4, the SPT. Most risk adverse strategies would likely select this tree, especially if there were a profound possibility that the demand intensity might exceed 10. Suppose, on the other hand, that some discrete probability distribution for β were available, e.g.,

β	2	4	6	8	10
Probability	0.2	0.4	0.4	0.1	0.1

The *minimum expected cost* design in this case, where higher values are less likely, would be T_2.

REFERENCES

Ahuja, R. K., Magnanti, T. L., & Orlin, J. B. (1993). *Network flows: Theory, algorithms, and applications*. New York: Prentice-Hall.

Annals of Operations Research. (2001). All articles. *106*.

Anandalingam, G., & Raghavan, S. (Eds). (2002). *Telecommunications network design and management*. Operations research/computer science interfaces series, 23. Dordrecht: Kluwer Academic Publishers.

Balakrishnan, A., Magnanti, T., Shulman, A., & Wong, R. (1991). Models for capacity expansion in local access telecommunications networks. *Annals of Operations Research*, *33*, 239–284.

Barnhart, C., Krishnan, N., Kim, D., & Ware, K. (2002). Network design for express shipment delivery. *Computational Optimization and Applications*, *21*, 239–262.

Gabow, H. N., & Myers, E. W. (1978). Finding all spanning trees of directed and undirected network. *SIAM Journal of Computing*, *7*, 280–287.

Hochbaum, D., & Segev, A. (1989). Analysis of a flow problem with fixed charges. *Networks*, *19*, 291–312.

Horowitz, E., Sahni, S., & Rajasekaran, S. (1997). *Computer algorithms*. San Francisco: W. H. Freeman Press.

Infager, G. (1994). *Planning under uncertainty*. Danvers, MA: Boyd & Fraser.

Khuller, S., Raghavachari, B., & Young, N. (1995). Balancing minimum spanning trees and shortest-path trees. *Algorithmica*, *14*, 305–321.

Magnanti, T., Wolsey, L., & Wong, R. (1992). *Network design. Handbooks in operations research and management science, vol. 6: Networks*. Amsterdam: North-Holland.

Matsui, T. (1997). A flexible algorithm for finding all the spanning trees in undirected networks. *Algorithmica*, *18*, 530–544.

Minoux, M. (1989). Network synthesis and optimum network design problems: Models, solution methods and applications. *Networks*, *19*, 313–360.

Render, B., & Stair, R. (1991). *Quantitative analysis for management* (4th ed.). Boston: Allyn and Bacon.

Riis, M., & Andersen, K. A. (2002). Capacitated network design with uncertain demand. *INFORMS Journal on Computing*, *14*, 247–260.

Sen, S., Doverspike, R., & Cosares, S. (1994). Network planning with random demands. *Telecommunications Systems*, *3*, 11–30.

Sherali, H., Subramanian, S., & Loganathan, G. (2001). Effective relaxations and partitioning schemes for solving water distribution network design problems to global optimality. *Journal of Global Optimization*, *19*, 1–26.

Tarjan, R. (1983). *Data structures and network algorithms*. CBMS-NSF Regional Conference Series in Applied Mathematics 44. Philadelphia: SIAM.

Vasko, F. J., Barbieri, R. S., Rieksts, B. Q., Reitmeyer, K. L., & Stott, K. L. (2002). The cable trench problem: Combining the shortest path and minimum spanning tree problems. *Computers and Operations Research*, *29*(5), 441–458.

APPENDIX. FINDING THE OPTIMAL SPECTRUM

As the spanning trees in the network are enumerated, the values of $F(T)$ and $V(T)$ for each are calculated and the information is inserted into a binary heap, which organizes the trees in increasing order of $F(T)$. The complete set of calculations require $O(kn \log n)$ time, while the heap is constructed in $O(k \log k)$ time (see Horowitz, et al., 1997). We assume there are no ties in the $F(T)$ values in the heap since ties can be resolved with the dominated tree removed during this construction phase. When completed, the top of the heap corresponds to MST which we denote T_1. Let $\beta_1 = 0$, denote the first confirmed value for $\phi*(\beta)$, and remove T_1 from the heap. Next, T_1 and β_1 are pushed onto a stack that stores trees that are believed to be optimal for some range of β. Some trees will be removed from the stack if they are later ruled out.

It follows from Theorem 1 that any tree in the heap having $V(T) > V(T_1)$ is dominated by T_1, so are successively removed from the top of the heap. If this empties the heap, then $T_1 = \text{MST} = \text{SPT}$, and $N = 1$, and we are done. Otherwise, set $i = 1$, and repeat the following steps:

Step 1. Let T be the tree at the top of the heap, and T_i be the tree at the top of the stack. It follows that $F(T) > F(T_i)$ and $V(T) < V(T_i)$. Calculate βT, which is the value of β where the cost functions of the two trees intersect.

Step 2. If $\beta T < \beta_i$, then the cost function for T_i is dominated by the cost function for T for the relevant range of β. Hence T_i is ruled out as an optimal tree, so:

- Pop (remove) T_i from the stack
- Decrease i by 1
- Go to Step 1

Otherwise T_i remains. Since T is also potentially optimal, it is placed onto the stack thus:

- Increase i by 1
- Set $T_i = T$
- Set β_i to βT
- Push (insert) T_i and β_i onto stack
- Go to Step 3

Step 3. Remove T from the heap. Successively remove from the top of the heap any tree S such that $V(S) > V(T)$. Such trees are dominated by T.

Step 4. If the heap is empty, we encounter $\text{SPT} = T_N$, so STOP. Otherwise go back to Step 1.

After this process is complete, the stack will contain all of the optimal trees with their associated breakpoints. Note that removing the top element from the heap and readjusting requires $O(\log k)$ time; pushing an element or popping the top element from a stack takes constant time. Each of the k trees is removed from the top of the heap once. So the total time to maintain the heap is $O(k \log k)$. The comparisons between the trees in the heap and the stack either result in a push or a pop operation. Note that no element in the stack is pushed or popped more than once, so the total time to maintain the stack is $O(k)$. Hence the entire process takes $O(k \log k)$ time.

This page intentionally left blank

A NOVEL APPROACH TO THE CONTINUOUS FLOW TRUCKLOAD ROUTING PROBLEM

Virginia M. Miori

ABSTRACT

Truckload routing has always been a challenge. This paper explores the development of continuous flow truckload routes, which resemble less than truckload routes, and a new way to formulate the truckload routing problem (TRP). Rather than view the problem as a succession of origin/destination pairs, we look at the problem as a series of routing triplets. This enables us to use alternate solution methods, which may result in greater efficiency and improved solutions.

1. INTRODUCTION

The logistics industry has always provided an abundance of challenges for the field of Operations Research. It has been a continuous source of problems in network design, facility layout, transportation mode selection, and vehicle routing and scheduling among other areas. As technology has

Applications of Management Science: In Productivity, Finance, and Operations
Applications of Management Science, Volume 12, 145–155

ISSN: 0276-8976/doi:10.1016/S0276-8976(06)12011-8

progressed, even more opportunities for improvement have been generated. This paper focuses specifically on vehicle routing and scheduling, more specifically the truckload routing problem (TRP).

The importance of the TRP, its extensions and its restrictions can be easily demonstrated by examining our national truckload transportation bill. The term truckload transportation extends an umbrella over various segments of the transportation industry, including but not limited to private fleet operations, truckload common carriers, truckload contract carriers, and container movement fleets. According to the Bureau of Transportation Statistics, truckload freight accounted for over $300 million of the total GDP (approximately $10 trillion) in the year 2001 (Bureau of Transportation Statistics). It is apparent that even a small improvement in truckload costs would garner impressive reductions in the overall transportation bill.

2. TRUCKLOAD ROUTING PROBLEM

Truckload operations, no matter the type industry supported, all share several common characteristics. The vehicle (or container) is loaded once with freight that either meets the cubic capacity or weight capacity of the equipment. All freight loaded is taken to a single destination and completely unloaded at that location.

The practice of truckload routing traditionally begins with a set of nodes which represents all possible lanes (origins/destination pairs) (Desrochers, Lenstra, Savelsbergh, & Soumis, 1988). The nodes are logically viewed in pairs or vehicle movements, one node being categorized as an origin and the other categorized as a destination. The base distribution scenario calls for a loaded movement to be followed by an empty return to the vehicle domicile. In order to defray the cost of the empty return movement, truckload carriers will attempt to pick up an additional truckload on the way back to the domicile. Once the vehicle is back at its domicile, the load assignment process is repeated.

The continuous flow (TRP) attempts to extend this traditional method of operation by building extended routes, which combine origin/destination pairs. Rather than returning to its domicile, a vehicle will instead proceed on to pick-up and deliver a subsequent load, which may take it even farther from the domicile. A vehicle stays out on the road for as long as allowable by the department of transportation (DOT) hours of service (HOS) rules. For the remainder of this paper, we will use TRP to represent the continuous flow version of the problem.

The TRP has the advantage of creating consecutive loaded movements while maintaining consistency in the route building approach. The movements are chained together into route sequences, which typically alternate between loaded and empty movements. The final routes that are created by the TRP may generally be performed at a lower cost than traditional operations and, when displayed on a map, resemble less-than-truckload (LTL) petal routes. We must employ alternate methods than the LTL problem due to many factors, a few of which include: geographic characteristics, nature of time windows, and cost characteristics.

This paper supports the emphasis of the TRP on continuous flow routing while providing an alternate approach to the formulation of the TRP. Rather than creating origin/destination pairs, we create node triplets. A node triplet is composed of two origin/destination pairs in succession and may be made up of any two loaded and/or empty movements. The final node of a triplet must match the subsequent triplet origin (first node). This method has been chosen to more effectively consider the opportunity cost of each movement.

A second important aspect of this formulation is that it may apply to either deterministic or stochastic TRP. We look specifically at the deterministic application in this paper, but emphasize that the stochastic formulation may be effectively generated by simply replacing known cost and demand data with probability distributions. The stochastic representation of the problem also requires alterations in the solution method, which will not be addressed in this paper.

The costs associated with the loaded and empty movements vary in accordance with explicit costs such as mileage between the nodes, and both nodes' geographic locations, as well as opportunity costs which result from overall demand patterns and loads foregone at a specific point in time. We clearly need an effective determination of these costs as applied to each load.

3. BACKGROUND

The traveling salesman problem (TSP) has evolved into one of the most studied problems in Operations Research and Computer Science literature. It is also well known that the TSP seeks an optimal solution to a single tour serving *n* customers. The probabilistic traveling salesman problem (PTSP) is presented as a variation on the TSP. Instead of a single tour to be performed once, the PTSP assumes repeated use of a tour, with each instance of the tour requiring a different subset of the customers to be visited. Rather than re-optimizing a TSP for each subset of customers, a single tour of all customers is established. The same tour is followed for each subset of customers, simply eliminating stops at those customers that need not be visited during that instance.

The TSP may be stated more quantitatively as a creation of a single routing tour of n points. The PTSP may then be restated as a creation of a single routing tour that serves a subset k of n points where $0 \leqslant k \leqslant n$. A probability distribution is used to specify the number and identity of k points to be visited in a given random instance of the problem.

The objective of the PTSP is to find the expected value of the minimum length tour. When solving the PTSP, we must first have a "good" a priori tour through all n points. This tour must also be good when only k points are present in a particular instance of the problem.

The PTSP is ideally suited as a basis for the truckload routing problem (TRP). Like the PTSP, the TRP requires a subset of all customers to be visited on a regular basis. The subset is known a priori on a given day, but is subject to variation and uncertainty in designation of future movements. Owing to the additional constraints, which we will impose on the PTSP, the TTP may be seen as a restriction of the PTSP.

4. PREVIOUS SOLUTION METHODS

Many solution methods have been used to solve problems of this nature. These include savings-based heuristics (Gronalt, Hartl, & Reimann, 2003), where a relaxed problem formulation, based on network flows, is used to calculate a lower bound to the solution value. Successive linear approximation heuristics (Frantzaskakis & Powell, 1990) employ rolling horizon procedures to simulate the operation of a truckload carrier.

A new branch and bound approach for solving an integer-programming formulation of the vehicle routing problem (VRP) with full truckloads is proposed (Arunapuram, Mathur, & Solow, 2003). It also considers time windows and waiting costs. The column generation scheme exploits the structure of the problem to solve the linear-programming relaxation problems that arise at the nodes. A two-path cut to produce better lower bounds and develop an effective separation algorithm allows the incorporation of inequalities into the master problem of a Dantzig-Wolfe decomposition approach (Kohl, Desrosiers, Madsen, Solomon, & Soumis, 1999). The sub-problems are shortest path problems with time window and capacity constraints. Branch and bound is applied to obtain integer solutions to the sub-problems.

5. TRP TRIPLET FORMULATION

We have already noted that the TRP is a restriction of the PTSP. Our discussion now moves to the assumptions and conditions of the TRP since these determine the constraints in the model formulation. The most basic assumption of the TRP is that freight is delivered by truck, in full truckload quantities. This means that a single unit of demand may not be split between multiple vehicles.

We now consider the definition of a route. It is composed of a series of triplets to be serviced by a single vehicle. We must begin and end a route at the same node, known as a domicile. Multiple domiciles may be present in a single instance of the TRP model, though this paper addresses only one.

Time conditions are placed on the handling of freight by the customers. Loads must be picked up and/or delivered on particular days of the week during specified times of the day. Some customers will be more restrictive than others in specifying their time windows.

The remaining conditions are imposed by the DOT. Restrictions are placed on the number of driving hours in a given week and number of work hours in a given week.

As previously stated we utilize the concept of a triplet. Recall that it may be composed of a loaded and empty movement, two loaded movements or two empty movements. The cost of the triplet is therefore the sum of the costs of the individual movements. The transit and work times are the sums of the transit and work times of the individual movements.

The TRP can be stated as a restricted PSTP. We customize the standard TSP constraints for our triplet formulation and supplement with load, schedule and triplet feasibility, vehicle utilization, and DOT regulations.

The following notation is required for the TRP:

N	the set of all nodes 1, … , N
V	the set of all vehicles 1, … , V
ijk	the set of all triplets $i,j,k \in 1, \ldots, N$
C_{ijk}	cost to serve triplet ijk
x_{ijk}^{v}	1 if triplet ijk is served by vehicle v 0 otherwise
y_{ij}	1 if node pair ij represents a loaded movement 0 otherwise
D_i	departure time from node i
t_{ij}	travel time between nodes i and j
$[e_i, l_i]$	the early and late arrival time window for node i

H	total allowed transit hours per week
w_i	work time for node i
W	total allowed work hours per week

The objective function may be stated as a cost minimization or a profit maximization. The financial structure of the transportation provider will dictate which of these approaches is preferred. It is likely that a private fleet would choose cost minimization, while a for-hire carrier would select profit maximization. In this paper the formulation is presented as a cost minimization.

$$\text{Minimize} \sum_i \sum_j \sum_k c_{ijk} x^v_{ijk}$$

A routing problem always has specific characteristics that must be modeled, and as such the constraints may be easily categorized into sets. The first set of constraints guarantees that each lane pair with demand (available load) is serviced at least once. A lane may be served as the first leg of a triplet (origin node to intermediate node) or as the second leg of the triplet (intermediate node to destination node).

$$\sum_v \sum_k y_{ij}(x^v_{ijk} + x^v_{kij}) = 1, \quad \forall y_{ij} = 1$$

The load must be carried and additional empty movements may also utilize the same lane. If a lane has no demand, it need not be used in the overall solution. The nature of the optimization ensures that lanes with no demand would only be used and/or reused for empty movements in support of a minimum cost or maximum profit solution. Since we employ a triplet formulation, we must combine lane level information (demand) with triplet level information (decision variables representing inclusion or exclusion of a triplet in the solution) in this constraint set.

The next constraint sets preserves traditional conservation of flow and schedule feasibility. The conservation of flow constraints ensure that every vehicle departs from the final node of every triplet they serve.

$$\sum_{ijlmv} (x^v_{ijk} - x^v_{kl}) = 0, \quad \forall k$$

The schedule feasibility constraints ensure a logical progression through the routes (a vehicle may not depart prior to its arrival), and also forces adherence to time constraints (time windows are satisfied for the pickup and delivery points as well as the domicile). Owing to the fact that time windows address individual locations, the time window constraints reflect the individual nodes and not the entire triplet. These are standard constraints used in the TRP.

$$D_i + t_{ij} \leq D_j, \quad \forall i,j$$

$$e_i \leq D_i \leq l_i, \quad \forall i$$

The remaining constraints set ensures that the routes satisfy the DOT regulations for single drivers. Each route is restricted to a maximum number of hours in transit and a maximum number of hours of work. The DOT recently revised these regulations, but the new regulations have been challenged. The model discussed in this paper allows for flexibility in the statement of these constraints to reflect this uncertainty.

$$\sum_{ijk} (t_{ij} + t_{jk}) x^v_{ijk} \leq H, \quad \forall v$$

$$\sum_{ij} (w_i + w_j) y_{ij} \leq W, \quad \forall v$$

6. COMPLETE DETERMINISTIC FORMULATION

$$\text{Minimize} \sum_i \sum_j \sum_k c_{ijk} x^v_{ijk}$$

subject to

$$\sum_v \sum_k y_{ij} (x^v_{ijk} + x^v_{kij}) = 1, \quad \forall y_{ij} = 1$$

$$\sum_{ijlmv} x^v_{ijk} - x^v_{kl} = 0, \quad \forall k$$

$$D_i + t_{ij} \leq D_j, \quad \forall i,j$$

$$e_i \leq D_i \leq l_i, \quad \forall i$$

$$\sum_{ijk} (t_{ij} + t_{jk}) x^v_{ijk} \leq H, \quad \forall v$$

$$\sum_{ij}(w_i + w_j)y_{ij} \leq W, \quad \forall v$$

$$x_{ijk}^v \in \{0, 1\}$$

7. COST SPECIFICATION

Direct and indirect costs make up the cost to move truckload freight over a lane. The direct costs include the fixed and variable costs of the operation. These costs are easily quantifiable because they draw directly from the fleet operation costs. Everything from ownership of vehicles to maintenance, fuels costs and taxes are included.

The indirect costs are much more difficult to quantify. Opportunity cost falls into this category. Carrier must try to determine the potential for increased access to demand, which might result from alternate routing decisions. They must also try to determine how much additional cost is associated with empty movements that might have been avoided had alternate decisions been made.

Another element that is interwoven in indirect costs is deterrence. Deterrence costs are an elusive yet important component of cost. They reflect the balance of freight availability within the transportation network. Nodes may be viewed as either a supply center[1] or a demand center[2]. A transportation company is typically willing to move freight at a tighter profit margin when moving to a supply center and seek a greater profit when moving to a demand center.

The direct costs include one predictable element, the actual cost of moving the load. The remaining costs: empty cost, opportunity cost, and deterrence cost, will vary based on the loads available and the overall freight demand patterns for the truckload industry.

Empty costs, which represent the cost of moving an unloaded vehicle from a load destination to the next possible load origin, must be applied to a loaded movement in the event of a profit maximization problem. A simple resolution to the empty cost allocation is to determine the empty cost per mile for each lane, find the expected number of empty miles for each lane and apply the total empty cost to the preceding loaded lane.

The final cost per loaded movement is usually specified on a per mile basis for each lane and includes all direct and indirect costs associated with that lane.

8. SAMPLE PROBLEM

The triplet formulation represents a departure from the tradition lane formulation of the TRP. This departure provides an opportunity for the application of new solution methods. It is therefore important to validate the triplet formulation against the lane formulation. We use a small sample problem to perform the initial validation.

The sample problem examines 5 nodes and 10 vehicles (Fig. 1). The model data includes the cost of a movement between each pair of nodes, the distance between each pair of nodes and the load requirements. Seventeen loads must be scheduled. Note that the number of vehicles may exceed the actual number required for the solution. The determination of the fleet size may be made from the solution of the TRP.

The sample problem was solved using both the triplet formulation and the lane formulation of the TRP (Fig. 2). The time window constraints were relaxed in order to test the basic validity of the triplet model formulation. The solutions were then generated using mixed integer (binary) linear programs. These LPs were written in AMPL and solved using the NEOS Server for Optimization. Both formulations resulted in the same optimal solution of $18495.28.

9. CONCLUSION

The TRP model is traditionally formatted with decision variables that represent freight carriage on particular lanes by particular vehicles. We discussed an alternate formulation, which used triplets rather than lanes. The triplet formulation produced a solution with the same optimal value as the

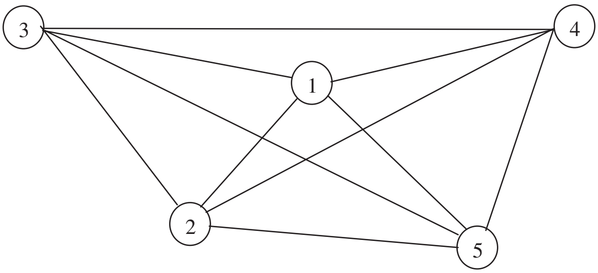


Fig. 1. Sample Problem Node Network

Vehicle Number	Triplet Formulation Triplets Scheduled	Lane Formulation Lanes Scheduled
1	232 - 224	21 – 13 - 35
2		34 – 41 - 45
3	515	
4	445 - 552	12 – 52 - 53
5	313 - 414	51
6	121 - 454	
7	353	14 – 15 – 23 - 31
8		32 - 41
9	434	43
10		24 - 45

Fig. 2. Solutions to the TRP

lane formulation. We may therefore conclude that the triplet formulation is a valid model.

The continuing research will examine alternate solution approaches, which exploit the unique structure of the triplet formulation. We may also extend the model from a deterministic formulation (as presented in this paper) to a stochastic formulation through the specification of probability distribution functions to represent the uncertain demand.

Cost structures also contain a high level of uncertainty in reality. Rather than using cost allocations made up entirely of easily quantifiable direct costs, a stochastic model would allow for representation of indirect costs and intangible costs though the use of probability distribution functions as well. The incorporation of both of these probabilities into the TRP results in a stochastic TRP.

NOTES

[1] A freight imbalance, which reflects a greater number of loads departing from a node than those arriving at a node.

[2] A freight imbalance, which reflects a greater number of loads arriving at a node than those departing from a node.

REFERENCES

Arunapuram, S., Mathur, K., & Solow, D. (2003). Vehicle routing and scheduling with full truckloads. *Transportation Science*, *37*(2), 170–182.

Bureau of Transportation Statistics, www.bts.com.

Desrochers, M., Lenstra, J. K., Savelsbergh, M. W. P., & Soumis, F. (1988). *Vehicle routing with time windows: Optimization and approximation*. Vehicle Routing: Methods and Studies, Elsevier: North Holland.

Gronalt, M., Hartl, R. F., & Reimann, M. (2003). New savings based algorithms for time constrained pickup and delivery of full truckloads. *European Journal of Operations Research*, *151*, 520–535.

Frantzaskakis, L. F., & Powell, W. B. (1990). A successive linear approximation procedure for stochastic, dynamic vehicle allocation problems. *Transportation Science*, *24*(1), 40–57.

Kohl, N., Desrosiers, J., Madsen, O. B. G., Solomon, M., & Soumis, F. (1999). 2-Path cuts for the vehicle routing problem with time windows. *Transportation Science*, *33*(1, Feb), 101–116.

NEOS Server for Optimization, www-neos.mcs.anl.gov/neos/.

This page intentionally left blank

VERTICAL COOPERATIVE ADVERTISING IN A MANUFACTURER–RETAILER SUPPLY CHANNEL

Susan X. Li, Zhimin Huang and Allan Ashley

ABSTRACT

Recent market structure reviews have shown a shift of retailing power from manufacturers to retailers. Retailers have equal or even greater power than a manufacturer when it comes to retailing. Based on this new market phenomenon, we intend to explore the role of vertical cooperative (co-op) advertising with respect to transactions between a manufacturer and a retailer. In this paper, we explore the role of vertical co-op advertising efficiency of transactions between a manufacturer and a retailer. We address the impact of brand name investments, local advertising, and sharing policy on co-op advertising programs in a manufacturer–retailer supply chain. Game theory concepts form the foundation for the analysis. We begin with the classical co-op advertising model where the manufacturer, as the leader, first specifies its strategy. The retailer, as the follower, then decides on its decision. We then relax the assumption of retailer's inability to influence the manufacturer's decisions and discuss full coordination between the manufacturer and the retailer on co-op advertising.

Applications of Management Science: In Productivity, Finance, and Operations
Applications of Management Science, Volume 12, 157–173

ISSN: 0276-8976/doi:10.1016/S0276-8976(06)12012-X

1. INTRODUCTION

Vertical cooperative (co-op) advertising is an interactive relationship between a manufacturer and a retailer in which the retailer initiates and implements a local advertisement and the manufacturer pays part of the costs. It is often used in consumer goods industries and plays a significant role in market strategy for many companies. In 1970, estimated co-op advertising expenditures spent by U.S. companies are up to $3 billion (Wolfe & Twedt, 1974). In 1980, it was estimated that approximately $5 billion was used in co-op advertising (Advertising Age, 1981), a 67% increase compared with 1970. More recently, about $30 billion was used in 1998 for co-op advertising (Davis, 1999). The continuous increase of spending volume and the growing importance of co-op advertising motivated us to explore the role of co-op advertising coordination and effective transactions in a manufacturer–retailer supply chain.

The main reason for a manufacturer to use co-op advertising is to motivate immediate sales at the retail level (Hutchins, 1953). The manufacturer's national advertising is intended to influence potential consumers to consider its brand and to help develop brand knowledge and preference, and is also likely to yield benefits beyond sales from an individual retailer. Retailer's local advertising gets people into the store and, with the passage of time, brings potential consumers to the stage of desire and action and gives an immediate reason to buy (brands being offered, specific prices, store location, etc.). Co-op advertising provides consumers the information needs when they move through the final stages of purchase and a congruence of information and information needs that would be impossible if the manufacturer uses only national advertising (Young & Greyser, 1983). In addition to the same objective of immediate sales at the retail level as the manufacturer, the retailer utilizes co-op advertising to reduce substantially its total promotional expense by sharing the cost of advertising with the manufacturer.

Most studies to date on vertical co-op advertising have focused on a relationship, where the manufacturer is a leader and the retailer is a follower, which implies that the manufacturer dominates the retailer. The design and management is the main subject (see, for example, Crimmins, 1970, 1985; Berger, 1972; Fulop, 1988; Hutchins, 1953; Somers, Gupta, & Herriott, 1990; Young & Greyser, 1983). Little attention has been given to the recent market structure in which retailers retain equal or more power than manufacturers do in retailing (Huang, Li, & Mahajan, 2002). This paper is intended to discuss the relationship between co-op advertising and the efficiency of manufacturer–retailer transactions. The results lead to the development of game theory structure that enables us to examine the problem of coordination in co-op advertising. Focusing on coordinately organizational economics between the manufacturer and the retailer differentiates this research from previous studies in the literature. In order to avoid the distraction of multiple products, multiple manufacturers, or multiple retailers, a system composed of a single manufacturer and a single retailer is selected to investigate the basic efficiency issue of coordination. Once the methodology is developed and basic issues are initially investigated, further research can be pursued to generalize the model to include multiple manufacturers and/or retailers.

In Section 2, we begin by delineating the assumed relationship and decision variables of the manufacturer and the retailer. The sales response function of the product at the retail level is assumed explicitly nonlinear in the manufacturer's brand name investments and the retailer's local advertising level, which is different from the literature where the retailer's local advertising level is the only factor.

Section 3 formulates the relationship between the manufacturer and the retailer as a classical ''leader–follower'' two-stage game. In this classical co-op advertising structure, the manufacturer, as the leader, first specifies the brand name investments and the co-op reimbursement policy. The retailer, as the follower, then decides on the local advertising level. The Stackelberg equilibrium is achieved.

We address our analysis of fully coordinated co-op advertising in Section 4. We relax the leader–follower structure by assuming a symmetric relationship between the manufacture and the retailer. We focus on the discussion of the transactions–efficiency for co-op advertising. We show that, (i) all Pareto efficient co-op advertising schemes are associated with a single local advertising level and a single brand name investment quantity; (ii) among all possible co-op advertising schemes, the system profit (the sum of the manufacturer's and the retailer's profits) is maximized for every Pareto efficient scheme, but not for any other schemes; (iii) the system profit at any Pareto efficient scheme is higher than at Stackelberg equilibrium; (iv) the manufacturer's brand name investments at full coordination is higher than at Stackelberg equilibrium; (v) the local advertising level at full coordination is higher than at Stackelberg equilibrium; and (vi) there is a subset of Pareto efficient co-op advertising schemes on which both the manufacturer and the retailer achieve higher profits than at Stackelberg equilibrium and which are determined by the sharing policy of the local advertising expenditures between the manufacturer and the retailer.

Among those feasible Pareto efficient co-op advertising schemes, the question is which one is the best (sharing policy) for both system members. We address this issue and consider the Nash bargaining model for determining the best sharing policy. The Nash model predicts that both the manufacturer and the retailer should equally share the system additional profits.

Concluding remarks are in Section 5. All proofs of results are in the appendix.

2. ASSUMPTIONS

(i) The retailer's sales response volume function of the product, S, is assumed to be affected mainly by the retailer's local advertising level, a, and the manufacturer's national brand name investments, q, which include national advertising and control of implementing co-op advertising agreement between the manufacturer and the retailer. As Young and Greyser (1983) point out that co-op advertising is used to attract the attention of customers near the time of actual purchase and therefore it is to stimulate short-term sales. The manufacturer's brand name investments such as the national advertising is intended to take the potential customers from the awareness of the product to the purchase consideration. The function of the local advertising is to bring potential customers to the stage of desire and action, to give reasons such as low price and high quality to buy, and to state when and where to obtain the product. Therefore, the manufacturer's brand name investments and the retailer's local advertising perform different but complementary functions that have positive effects on the ultimate product sales. Saturation may be reached when both or either the local advertising efforts and the brand name investments are increased. Since co-op advertising is intended to generate short-term sales, we may consider one-period sales response volume function as $S(a, q) = \alpha – \beta a^{-\gamma} q^{-\delta}$, where $\alpha > 0$ is the sales saturate asymptote and β, γ, and δ are positive constants. There is a substantial literature on the estimation of sales response volume functions (see, for example, Little, 1979), but all of them consider only the local advertising effect, not others such as national advertising on the volume of sales.

(ii) The manufacturer's dollar marginal profit for each unit to be sold is ρ_{m}, and the retailer's dollar marginal profit is ρ_{r}.

(iii) The fraction of total local advertising expenditures, which manufacturer agrees to share with retailer is t, which is the manufacturer's co-op advertising reimbursement policy.

(iv) The manufacturer's, retailer's, and system's profit functions are as the following:

$$\pi_{\mathrm{m}} = \rho_{\mathrm{m}}(\alpha - \beta a^{-\gamma} q^{-\delta}) - ta - q \tag{1}$$

$$\pi_{\mathrm{r}} = \rho_{\mathrm{r}}(\alpha - \beta a^{-\gamma} q^{-\delta}) - (1 - t)a \tag{2}$$

$$\pi = \pi_{\mathrm{m}} + \pi_{\mathrm{r}} = (\rho_{\mathrm{m}} + \rho_{\mathrm{r}})(\alpha - \beta a^{-\gamma} q^{-\delta}) - a - q \tag{3}$$

3. CLASSICAL CO-OP ADVERTISING MODEL

In this section, we model the relationship between the manufacturer and the retailer, as a two-stage game with the manufacturer as the leader and the retailer as the follower. The solution of the game is called Stackelberg equilibrium. This relationship may be explained as follows. The original idea of co-op advertising came from the demands of the retailer's promotional help from the manufacturer in order to increase the retailer's advertising budgets without spending more of retailer's own funds. In the absence of the manufacturer's co-op advertising funds, the retailer will usually spend less money on the local advertising than the amount that is optimal from the manufacturer's point of view. The manufacturer can use co-op advertising subsidization policy to induce the retailer to increase its local advertising expenditure at a level that results in additional sales of the product to the retailer and, thereby, to the manufacturer. The determination of the level of local advertising expenditures depends on how much the manufacturer is willing to subsidize the retailer. The manufacturer may set up some requirements on the co-op advertising, such as the size of the advertisement, the display of the manufacturer's brand name, and certain product features. The manufacturer, as the leader, first declares the level of brand name investments and the co-op advertising policy. The retailer, as the follower, then decides on the quantity of products to be purchased from the manufacturer taking into account the total local advertising expenditures to be spent. In other words, the retailer takes the (Stackelberg) equilibrium local advertising expenditures into account in deciding the volume of product to be ordered. The manufacturer on the other hand maximizes its profits by specifying the level of brand investments and co-op advertising reimbursement taking the behavior of the retailer into account.

In order to determine Stackelberg equilibrium, we first solve the reaction functions in the second stage of the game. Since π_{r} is a concave function of a, the optimal value of the local advertising expenditures is determined by setting the first derivative of π_{r} with respect to a to be zero:

$$\frac{\partial \pi_{\mathrm{r}}}{\partial a} = \gamma \rho_{\mathrm{r}} \beta a^{-(\gamma+1)} q^{-\delta} - (1 - t) = 0 \tag{4}$$

Then, we have

$$a = \left(\frac{\gamma \rho_{\mathrm{r}} \beta}{(1 - t) q^{-\delta}} \right)^{1/(\gamma+1)} \tag{5}$$

Eq. (5) describes positively and negatively changes in responding to the changes in manufacturer's co-op advertising reimbursement policy and brand name investments. These can be seen by observing that

$$\frac{\partial a}{\partial t} = \frac{1}{\gamma + 1} (\gamma \rho_{\mathrm{r}} \beta / q^{\delta})^{1/(\gamma+1)} (1 - t)^{-(2+\gamma)(\gamma+1)} > 0 \tag{6}$$

$$\frac{\partial a}{\partial q} = -\frac{\delta}{\gamma + 1} (\gamma \beta \rho_{\mathrm{r}} / (1 - t))^{1/(\gamma+1)} q^{-(\gamma+\delta+1)(\gamma+1)} < 0 \tag{7}$$

Eq. (6) tells us that the more the manufacturer is willing to share the cost of local advertising, the more the retailer will spend on the local advertising. Therefore, the manufacturer's co-op advertising policy can be used as an indicator of the amount of money that the retailer would spend on local advertising. The manufacturer can use this indicator to induce the retailer to increase local advertising expenditure at a level that the manufacturer expects. It also shows that greater manufacturer's share of local advertising spending would lead to more retailer's spending for local advertising with the ultimate result of increased sales for both the retailer and the manufacturer. Eq. (7) tells us that in this leader–follower relationship, if the retailer is to capitalize effectively on the brand awareness created by the manufacturer's national advertising, he/she might to increase or decrease local advertising expenditures in accordance with the effect of national advertising. In other words, the retailer has strong incentive to spend less money on the local advertising if the manufacturer increases the level of brand name investments. The amount of money spent by the manufacturer on national advertising can be used as another indicator of the amount of money that the retailer would spend on local advertising.

An analysis of influence of the manufacturer's (leader) brand name investments and co-op advertising policy on the retailer's (follower) local advertising is very important. This analysis may allow the manufacturer to measure the effects of changes in manufacturer's brand name investments and co-op advertising policy from one period to another period retailer local advertising. The manufacturer can use it to determine: (a) when to do national advertising; (b) how long the national advertising would affect the retailer's local advertising efforts; (c) when to introduce new products; and (d) the total budgets of brand name investments and co-op advertising reimbursement.

Next, the optimal values of q and t are determined by maximizing the manufacturer's profit subject to the constraint imposed by (5). Hence, the manufacturer's problem can be formulated as

$$\begin{aligned}&\max_{q,\,t} \quad \pi_{\mathrm{m}} = \rho_{\mathrm{m}}(\alpha - \beta\alpha^{-\gamma}q^{-\delta}) - ta - q \\ &\text{s.t.} \\ &0 \le t \le 1, \;\; q \le 0 \end{aligned} \tag{8}$$

where $a = (\gamma\rho_{\mathrm{r}}\beta/(1-t)q^{\delta})^{1/(\gamma+1)}$.

Substituting $a = (\gamma\rho_{\mathrm{r}}\beta/(1-t)q^{\delta})^{1/(\gamma+1)}$ into the objective yields the following optimization problem for the manufacturer:

$$\begin{aligned}\text{Max} \quad \pi_{\mathrm{m}} = {} & \rho_{\mathrm{m}}[\alpha - \beta(\gamma\rho_{\mathrm{r}}\beta)^{-\gamma/(\gamma+1)}(1-t)^{\gamma/(\gamma+1)}q^{-\delta/(\gamma+1)}] \\ & - (\gamma\rho_{\mathrm{r}}\beta)^{1/(\gamma+1)}t(1-t)^{-1/(\gamma+1)}q^{-\delta/(\gamma+1)} - q \\ \text{s.t.} & \\ 0 \le t \le 1, & q \ge 0 \end{aligned} \tag{9}$$

Theorem 1. Let

$$a^* = [\delta^{-\delta}\beta\gamma^{\delta+1}(\rho_{\mathrm{m}} - \gamma\rho_{\mathrm{r}})]^{1/(\delta+\gamma+1)} \tag{10}$$

$$t^* = (\rho_{\mathrm{m}} - (\gamma+1)\rho_{\mathrm{r}})/(\rho_{\mathrm{m}} - \gamma\rho_{\mathrm{r}}) \tag{11}$$

$$q^* = [\delta^{\gamma+1}\beta\gamma^{-\gamma}(\rho_{\mathrm{m}} - \gamma\rho_{\mathrm{r}})]^{1/(\delta+\gamma+1)} \tag{12}$$

Then (a^*, t^*, q^*) is the equilibrium point of the two-stage game.

From the above optimal formulations, the fraction level t^* is positively and negatively correlated to changes in manufacturer's marginal profits and retailer's marginal profits, respectively. These are because of

$$\frac{\partial t^*}{\partial \rho_{\mathrm{m}}} = \frac{\rho_{\mathrm{r}}}{(\rho_{\mathrm{m}} - \gamma\rho_{\mathrm{r}})^2} > 0 \tag{13}$$

$$\frac{\partial t^*}{\partial \rho_{\mathrm{r}}} = \frac{-\rho_{\mathrm{m}}}{(\rho_{\mathrm{m}} - \gamma\rho_{\mathrm{r}})^2} < 0 \tag{14}$$

For manufacturer, if his marginal profit is high (for instance, those manufacturers who produce infrequently purchased goods such as appliances and linens), he/she knows that infrequently purchased products are not very standing out most noticeably to most consumers, except at the time of purchase or need. Once consumer decides to purchase this kind of product, one always or often makes an overt search among local sources of information, seeking specific product information. In order to give the retailer more incentive to attract consumers, the manufacturer should share more local advertising expenditures with the retailer. On the other, if retailer's marginal profit is high, at this situation retailer has strong incentive to spend money in local advertising to attract consumers to buy these products, even though the manufacturer only shares a small fraction of local advertising expenditures.

4. FULL COORDINATED VERTICAL CO-OP ADVERTISING MODEL

In previous section, we focused on the equilibrium results for a two-stage game structure. We assumed that the manufacturer as the leader holds extreme power and has almost complete control over the behavior of the retailer. The retailer is presumably powerless to influence the manufacturer. The relationship is that of an employer and an employee. The fact that, in many industries, manufacturing and retailing are vertically separated makes the effective implementation of the manufacturer's co-op advertising program difficult. The manufacturer who can best implement a national advertising campaign to promote brand awareness may not know the local market and the retailer's advertising behavior. The manufacturer can only provide such a co-op advertising program to the retailer and the retailer who knows the local market and can best advertise to create immediate sales, decides whether the offer is taken advantage of and to what extent. In other words, it is the retailer, not the manufacturer, decides how much, if any, of the manufacturer's money is spent. As Crimmins (1973) points out that approximately one-third to one-half of the manufacturer's money allocated through co-op advertising allowances is not used by its retailers.

Recent studies in marketing have demonstrated that in many industries, retailers have increased their power relative to manufacturers over the past two decades. The shift of power from manufacturers to retailers is one of the most significant phenomena in manufacturing and retailing. In consumer goods industries, one of the most influence of a retailer on the market performance of the manufacturer is the differentiation of the manufacturer's product. The retailer controls some of the attributes of the product, which the consumer may desire. The product's quality and image may be reflected by the retailer's store reputation and image. It is also true that the retailer can influence the sales of the manufacturer's product through local advertising and other selling efforts, such as a selling presentation, personal recommendation, or advice solicited by the consumer, which provide enough information about the product (Porter, 1974). The retailer transfers product information about reliability, features and method of use that may not be available from the manufacturer's national advertising efforts, and other sources. Especially, for durable goods such as appliances and automobiles, the retailer has more influence on the consumer's purchase decision. Although the manufacturer's national advertising can lead the consumer to consider a particular brand, the retailer's local advertising and sales efforts can be used to negate the effect of national advertising by changing the consumer's mind. The retailer is able to withhold its selling efforts which includes local advertising for a particular brand and to influence the consumer to purchase another brand. It is clear that as the retailer's influence on product differentiation increases, the retailer's bargaining power relative to the manufacturer increases. This enables the retailer to exercise its enhanced retailing power in order to extract local advertising allowances and additional discounts from the manufacturer. Many retailers actually use manufacturers' allowances and co-op advertising programs for their own purposes, and in process reduce their dependence on the manufacturers (see, for example, Achenbaum & Mitchel, 1987; Buzzell, Quelch, & Salmon, 1990; Fulop, 1988; Olver & Farris, 1989).

In this section, we relax the leader–follower relationship and assume a symmetric relationship between the manufacturer and the retailer. We will discuss the efficiency of manufacturer and retailer transactions in vertical co-op advertising agreements. Similar approaches have been used in distribution channels, franchising arrangements, and inventory control systems (Charnes, Huang, & Mahajan, 1995; Jeuland & Shugan, 1983; Kohli & Park, 1989; Li & Huang, 1995; Huang et al., 2002; Yue, Austin, Wang, & Huang, 2006).

Now, let us consider Pareto efficient advertising schemes in our co-op advertising arrangements. A scheme (a_0, t_0, q_0) is called Pareto efficient if one cannot find any other scheme (a, t, q) such that neither the manufacturer's nor the retailer's profit is less at (a, t, q), but at least one of the manufacturer's and retailer's profits is higher at (a, t, q) than at (a_0, t_0, q_0). More precisely, (a_0, t_0, q_0) is Pareto efficient if and only if $\pi_{\mathrm{m}}(a, t, q) \geq \pi_{\mathrm{m}}(a_0, t_0, q_0)$ and $\pi_{\mathrm{r}}(a, t, q) \geq \pi_{\mathrm{r}}(a_0, t_0, q_0)$ for some (a, t, q) implies that $\pi_{\mathrm{m}}(a, t, q) = \pi_{\mathrm{m}}(a_0, t_0, q_0)$ and $\pi_{\mathrm{r}}(a, t, q) = \pi_{\mathrm{r}}(a_0, t_0, q_0)$.

Since π_{m} and π_{r} are quasi-concave, the set of Pareto efficient schemes consists of those points where the manufacturer's and the retailer's iso-profit surfaces are tangent to each other, i.e.,

$$\nabla\pi_{\mathrm{m}}(a, t, q) + \mu\nabla\pi_{\mathrm{r}}(a, t, q) = 0 \tag{15}$$

for some $\mu \geq 0$ (see Charnes, Huang, Rousseau, & Wei, 1990), where $\nabla\pi_{\mathrm{m}} = \partial\pi_{\mathrm{m}}/\partial a, \partial\pi_{\mathrm{m}}/\partial t, \partial\pi_{\mathrm{m}}/\partial q$ stands for the gradient of π_{m}. This leads to the following theorem.

Theorem 2:. The collection of Pareto efficient schemes is described by the set

$$Y = \{(\overline{a}^*, t, \overline{q}^*) : 0 \leq t \leq 1\} \tag{16}$$

where -$\bar{a}^* = [\delta^{-\delta}\beta\gamma^{\delta+1}(\rho_{\mathrm{m}} + \rho_{\mathrm{r}})]^{1/(\delta+\gamma+1)}$ and $\bar{q}^* = [\delta^{\gamma+1}\beta\gamma^{-\gamma}(\rho_{\mathrm{m}} + \rho_{\mathrm{r}})]^{1/(\delta+\gamma+1)}$.

This theorem tells us that all Pareto efficient schemes are associated with a single local advertising expenditure $\bar{a}^*$ and a single manufacturer's brand name investment $\bar{q}^*$ and with the fraction t of the manufacturer's share of the local advertising expenditures between 0 and 1. The locus of tangency lies on a vertical line segment at $(\bar{a}^*, \bar{q}^*)$ in (a, t, q) space because the expressions for both iso-profit surfaces contain only linear fraction variable t, so that vertically shifting an iso-profit surface yields another iso-profit surface. When a pair of tangent iso-profit surfaces shift vertically, the tangent point also shifts vertically so that the locus of tangency traces a vertical line.

Theorem 3. An advertising scheme is Pareto efficient if and only if it is an optimal solution of

$$\begin{aligned} &\overline{\pi}^* \underset{a,\, t,\, q}{\mathrm{Max}} \quad \pi = \pi_{\mathrm{m}} + \pi_{\mathrm{r}} \\ &\text{s.t.} \\ &0 \leq t \leq 1, \;\; q \geq 0, \;\; a \geq 0 \end{aligned} \tag{17}$$

This theorem tells us that, among all possible advertising schemes, the system profit (i.e., the sum of the manufacturer's and the retailer's profits) is maximized for every Pareto efficient scheme, but not for any other schemes. The following theorem implies that Pareto efficiency yields (1) higher system profit than at Stackelberg equilibrium, (2) higher manufacturer's brand name investment than at Stackelberg equilibrium, and (3) higher local advertising expenditures than at Stackelberg equilibrium.

Theorem 4.

$$\bar{\pi}^* > \pi^*, \quad \bar{q}^* > q^*, \quad \bar{a}^* > a^* \tag{18}$$

where $-*$ and "*" represent coordination and Stackelberg equilibrium, respectively.

Theorem 4 leads to the possibility that both the manufacturer and the retailer can gain more profits compared with Stackelberg equilibrium. It should be noted that not all Pareto efficient schemes are feasible to both the manufacturer and the retailer. Neither the manufacturer nor the retailer would be willing to accept less profits at full coordination than at Stackelberg equilibrium. An advertising scheme $(\bar{a}^*, t, \bar{q}^*) \in Y$ is called feasible Pareto efficient, if

$$\Delta\pi_{\mathrm{m}}(t) = \pi_{\mathrm{m}}(\overline{a}^*,\ t,\ \overline{q}^*) - \pi_{\mathrm{m}}^* \geq 0 \tag{19}$$

and

$$\Delta\pi_{\mathrm{r}}(t) = \pi_{\mathrm{r}}(\overline{a}^*, t, \overline{q}^*) - \pi_{\mathrm{r}}^* \geq 0 \tag{20}$$

since only schemes satisfying (19) and (20) are acceptable for both the manufacturer and the retailer when they do coordinate. We then call

$$Z = \{(\overline{a}^*, t, \overline{q}^*) : \ \Delta\pi_{\mathrm{m}}(t) \geq 0,\ \ \Delta\pi_{\mathrm{r}}(t) \geq 0,\ \ (\overline{a}^*, t, \overline{q}^*) \in Y\} \tag{21}$$

the feasible Pareto efficient set of advertising schemes.

Let

$$k_1 = \beta\rho_{\mathrm{m}}[(a^*)^{-\gamma}(q^*)^{-\delta} - (\bar{a}^*)^{-\gamma}(\bar{q}^*)^{-\delta}] + (q^* - \bar{q}^*) + a^* t^* \tag{22}$$

$$k_2 = \beta\rho_{\mathrm{r}}[(a^*)^{-\gamma}(q^*)^{-\delta} - (\bar{a}^*)^{-\gamma}(\bar{q}^*)^{-\delta}] + (a^* - \overline{a}^*) - a^* t^* \tag{23}$$

$$t_{\min} = -k_2/\overline{a}^* \tag{24}$$

and

$$t_{\max} = k_1/\overline{a}^* \tag{25}$$

Here we suppose $k_2 < 0$. Then $\Delta\pi_{\mathrm{m}}(t) = k_1 - \overline{a}^* t, \quad \Delta\pi_{\mathrm{r}}(t) = k_2 + \overline{a}^* t$, and Z can be simplified as

$$Z = \{(\overline{a}^*, t, \overline{q}^*) : t_{\min} \le t \le t_{\max}\} \tag{26}$$

It can be shown that $1 > t_{\max} > t_{\min} \ge 0$ (see the appendix). Therefore, for any given t that satisfies $t_{\min} < t < t_{\max}$, $\Delta\pi_{\mathrm{m}}(t) > 0$ and $\Delta\pi_{\mathrm{r}}(t) > 0$. This simply implies that there exist Pareto efficient advertising schemes such that both the manufacturer and the retailer are better off at full coordination than at Stackelberg equilibrium. We are interested in finding an advertising scheme in Z, which will be agreeable to both the manufacturer and the retailer. According to Theorem 4, for any Pareto scheme $(\overline{a}^*, t, \overline{q}^*)$, $\Delta\pi_{\mathrm{m}}(\overline{a}^*, t, \overline{q}^*) + \Delta\pi_{\mathrm{r}}(\overline{a}^*, t, \overline{q}^*) = \Delta\pi$ where $\Delta\pi = \overline{\pi}^* - \pi^*$ is a positive constant.

We refer $\Delta\pi$ as the system profit gain since it is the joint profit gain achieved by the manufacturer and the retailer by moving from a Stackelberg advertising scheme to a Pareto efficient advertising scheme. This property implies that the more the manufacturer's share of the system profit gain, the less the retailer's share of the system profit gain, and vice versa. The property that all feasible efficient transactions occur at $(\overline{a}^*, \overline{q}^*)$ implies that the manufacturer and the retailer will agree to change the local advertising expenditures to $\overline{a}^*$ from a^* and the brand name investments to $\overline{q}^*$ from q^*. However, they will negotiate over the manufacturer's share of the local advertising expenditures t.

Assume that the manufacturer and the retailer agree to change local advertising expenditures to $\overline{a}^*$ and brand name investments to $\overline{q}^*$ from a^* and q^*, respectively, and engage in bargaining for the determination of reimbursement percentage to divide the system profit gain. A fraction closer to $t_{\max}$ is preferred by the retailer, and a fraction closer to $t_{\min}$ is preferred by the manufacturer. Let us utilize Nash (1950) bargaining model in our co-op advertising model. Since our entire problem is deterministic, we assume both the manufacturer and the retailer's utility functions, u_{m} and u_{r}, over $\Delta\pi_{\mathrm{m}}$ and $\Delta\pi_{\mathrm{r}}$ are linear. Without loss of generality, we assume $u_{\mathrm{m}}(\Delta\pi_{\mathrm{m}}) = \Delta\pi_{\mathrm{m}}$ and $u_{\mathrm{r}}(\Delta\pi_{\mathrm{r}}) = \Delta\pi_{\mathrm{r}}$.

According to Nash model, the best Pareto efficient advertising reimbursement, t, is obtained by the following problem:

$$\begin{aligned} &\text{Max} \quad \Delta\pi_{\mathrm{m}}(t)\Delta\pi_{\mathrm{r}}(t) \\ &\text{s.t.} \\ &t_{\min} \le t \le t_{\max} \end{aligned} \tag{27}$$

Since $\Delta\pi_{\mathrm{m}}(t)\Delta\pi_{\mathrm{r}}(t) = (k_1 - \overline{a}^{*}t)(k_2 + \overline{a}^{*}t)$, setting the first derivative of $\Delta\pi_{\mathrm{m}}(t)\Delta\pi_{\mathrm{r}}(t)$ to be zero, we have

$$\bar{t}^{*} = (t_{\min} + t_{\max})/2 \tag{28}$$

Substitute (28) into $\Delta\pi_{\mathrm{m}}(t)$ and $\Delta\pi_{\mathrm{r}}(t)$, we have

$$\Delta\pi_{\mathrm{m}}(\bar{t}^{*}) = \Delta\pi_{\mathrm{r}}(\bar{t}^{*}) = \Delta\pi/2 \tag{29}$$

Therefore, the Nash model predicts that both the manufacturer and the retailer should equally share the system additional profits.

5. CONCLUDING REMARKS

This paper attempts to investigate the efficiency of transactions for the system of manufacturer–retailer co-op advertising in the context of game theory. Two co-op advertising models are discussed and compared. The first is the classical leader–follower advertising model, where the manufacturer is the leader and the retailer is the follower. The second advertising model assumes the symmetric positions of the manufacturer and the retailer in the decision-making process and discusses efficient co-op advertising schemes. The Nash bargaining model is utilized to select the best co-op advertising expenditure-sharing rule between the manufacturer and the retailer.

There are three possible avenues for future research. First, the manufacturer–retailer relationship is multi-dimensional, including wholesale and retail price, credit terms, shelf-space, contributions toward special sales, treatment of competitors product, and policy with respect to house brands. It may hard, but should very interesting to take all or many of these factors into consideration in the future research. Second, the single manufacture–retailer system assumption can be relaxed to a duopoly situation of manufacturers who sell their products through a common monopolistic retailer and who sells multiple competing brands with varying degrees of substitutability. It would be interesting to discuss the impact of duopolistic manufacturer's brand name investments, monopolistic retailer's local advertising level, and advertising sharing rules among manufacturers and retailer on co-op advertising expenditures. Finally, in our analysis we employed nonlinear sales response function to satisfy the saturation requirement. As indicated in the literature of channel studies, many important results in equilibrium analyses depend on the shape of the product demand function

(Moorthy, 1988; Shugan, 1985). Therefore, the use of a linear sales response function may yield different and interesting results in the analysis for vertical co-op advertising agreements.

REFERENCES

Achenbaum, A. A., & Mitchel, F. K. (1987). Pulling away from push marketing. *Harvard Business Review*, *65*(May–June), 38–40.

Advertising Age. (1981). Partnership perks up profits. *Advertising Age*, August 17, p. S-1.

Berger, P. D. (1972). Vertical cooperative advertising ventures. *Journal of Marketing Research*, *9*, 309–312.

Buzzell, R. D., Quelch, J. A., & Salmon, W. J. (1990). The costly bargain of trade promotion. *Harvard Business Review*, *68*(March–April), 141–149.

Charnes, A., Huang, Z. M., & Mahajan, V. (1995). Franchising coordination with brand name considerations. *Research in Marketing*, *12*, 1–47.

Charnes, A., Huang, Z. M., Rousseau, J. J., & Wei, Q. L. (1990). Cone extremal solutions of multi-payoff games with cross-constrained strategy sets. *Optimization*, *21*, 51–69.

Crimmins, E. C. (1970). *A management guide to cooperative advertising*. New York: Association of National Advertisers.

Crimmins, E. C. (1973). A co-op myth: It is a tragedy that stores don't spend all their accruals. *Sales and Marketing Management*, (February 7), p. 7.

Crimmins, E. C. (1985). *Cooperative advertising*. New York: Gene Wolf & Co.

Davis, R. A. (1999). Retailers open doors wide for co-op. *Advertising Age*, *70*(August 1), 30.

Fulop, C. (1988). The role of advertising in the retail marketing mix. *International Journal of Advertising*, *7*, 99–117.

Huang, Z., Li, S. X., & Mahajan, V. (2002). An analysis of manufacturer–retailer supply chain coordination in cooperative advertising. *Decision Sciences*, *33*(3), 469–494.

Hutchins, M. S. (1953). *Cooperative advertising*. New York: Roland Press.

Jeuland, A. P., & Shugan, S. M. (1983). Managing channel profits. *Marketing Science*, *2*, 239–272.

Kohli, R., & Park, H. (1989). A cooperative game theory model of quantity discounts. *Management Science*, *35*, 693–707.

Li, X. S., & Huang, Z. M. (1995). Managing buyer–seller system cooperation with quantity discount considerations. *Computers and Operations Research*, *22*, 947–958.

Little, J. D. C. (1979). Aggregate advertising models: The state of the art. *Operations Research*, *27*, 627–629.

Moorthy, K. S. (1988). Strategic decentralization in channels. *Marketing Science*, *7*, 335–355.

Nash, J. F. (1950). The bargaining problem. *Econometrica*, *18*, 155–162.

Olver, J. M., & Farris, P. W. (1989). Push and pull: A one-two punch for packaged products. *Sloan Management Review*, *30*(Fall), 53–61.

Porter, M. E. (1974). Consumer behavior, retailer power and market performance in consumer goods industries. *Review of Economics and Statistics*, *LVI*, 419–436.

Shugan, S. (1985). Implicit understandings in channels of distribution. *Management Science*, *31*, 435–460.

Somers, T. M., Gupta, Y. P., & Herriott, S. R. (1990). Analysis of cooperative advertising expenditures: A transfer-function modeling approach. *Journal of Advertising Research*, *24*(October–November), 35–45.

Wolfe, H. D., & Twedt, D. W. (1974). *Essentials of the promotional mix*. New York: Appleton Century Crofts.

Young, R. F., & Greyser, S. A. (1983). *Managing cooperative advertising: A strategic approach*. New York: Lexington Books.

Yue, J., Austin, J., Wang, M., & Huang, Z. M. (2006). Coordination of cooperative advertising in a two-level supply chain when manufacturer offers discount. *European Journal of Operational Research*, *168*, 65–85.

APPENDIX. PROOF OF RESULTS

Proof of Theorem 1. Solving the first-order conditions of π_{m} with respect to t and q yields

$$t^* = \frac{\rho_{\mathrm{m}} - (1+\gamma)\rho_{\mathrm{r}}}{\rho_{\mathrm{m}} - \gamma\rho_{\mathrm{r}}} \tag{A1}$$

and

$$q^* = [\delta^{\gamma+1}\beta\gamma^{-\gamma}(\rho_m - \gamma\rho_r)]^{1/(\delta+\gamma+1)} \tag{A2}$$

Substitute (A1) and (A2) into (5), we have

$$a^* = [\delta^{-\delta}\beta\gamma^{\delta+1}(\rho_{\mathrm{m}} - \gamma\rho_{\mathrm{r}})^{1/(\delta+\gamma+1)} \tag{A3}$$

Hence (a^*, t^*, q^*) is the optimal solution of (9).

Proof of Theorem 2. Since

$$\nabla\pi_{\mathrm{m}}(a,t,q) = (\beta\gamma\rho_{\mathrm{m}}a^{-(\gamma+1)}q^{-\delta} - t, -a, \beta\delta\rho_{\mathrm{m}}a^{-\gamma}q^{-(\delta+1)} - 1) \tag{A4}$$

and

$$\nabla\pi_{\mathrm{r}}(a,t,q) = (\beta\gamma\rho_{\mathrm{r}}a^{-(\gamma+1)}q^{-\delta} - (1-t), -a, \beta\delta\rho_{\mathrm{r}}a^{-\gamma}q^{-(\delta+1)} - 1) \tag{A5}$$

utilizing (15) we can get $\mu = 1$, $\bar{a}^*$ and $\bar{q}^*$ in (16) and with t between 0 and 1.

Proof of Theorem 3. Since $\pi = (\rho_{\mathrm{m}} + \rho_{\mathrm{r}})(\alpha - \beta a^{-\gamma}q^{-\delta}) - a - q$ does not contain the variable t, any value of t between 0 and 1 can be a component for any optimal solution.

Taking the first derivatives of π with respect to a and q, and setting them to 0, we have $\bar{a}^* = [\delta^{-\delta}\beta\gamma^{\delta+1}(\rho_{\mathrm{m}} + \rho_{\mathrm{r}})]^{1/(\delta+\gamma+1)}$ and $\bar{q}^* = [\delta^{\gamma+1}\beta\gamma^{-\gamma}(\rho_{\mathrm{m}} + \rho_{\mathrm{r}})]^{1/(\delta+\gamma+1)}$. Therefore, $(\bar{a}^*, t, \bar{q}^*)$ for any t in [0, 1] is an optimal solution of (17).

Proof of Theorem 4. (i) Proof of $\bar{\pi}^* > \pi^*$:Let

$$f(x) = (\gamma + \delta + 1)x + (\gamma + 1)\rho_{\mathrm{r}}/(\rho_{\mathrm{m}} + \rho_{\mathrm{r}}) - (\gamma + \delta + 1)x^{(\gamma+1)/(\gamma+\delta+1)}.$$

Since

$$\begin{aligned}\pi^* = {} & \alpha(\rho_{\mathrm{m}} + \rho_{\mathrm{r}}) - [(\gamma + \delta + 1)(\rho_{\mathrm{m}} - \gamma\rho_{\mathrm{r}}) + (\gamma + 1)\rho_{\mathrm{r}}] \\ & \times[(1/\delta)^{\delta}\beta\gamma^{-\gamma}(\rho_{\mathrm{m}} - \gamma\rho_{\mathrm{r}})^{-(\gamma+\delta)}]^{1/(\gamma+\delta+1)}\end{aligned}$$

and

$$\pi^* = \alpha(\rho_{\mathrm{m}} + \rho_{\mathrm{r}}) - (\delta + \gamma + 1)(\rho_{\mathrm{m}} + \rho_{\mathrm{r}})[(1/\delta)^{\delta}\beta\gamma^{-\gamma}(\rho_{\mathrm{m}} + \rho_{\mathrm{r}})^{-(\gamma+\delta)}]^{1/(\delta+\gamma+1)},$$

we have

$$\begin{aligned}\bar{\pi}^* - \pi^* = {} & (\rho_{\mathrm{m}} + \rho_{\mathrm{r}})(\rho_{\mathrm{m}} - \gamma\rho_{\mathrm{r}})^{-(\gamma+\delta)/(\gamma+\delta+1)}[(1/\delta)^{\delta}\beta\gamma^{-\gamma}]^{1/(\gamma+\delta+1)} \\ & \times f((\rho_{\mathrm{m}} - \gamma\rho_{\mathrm{r}})/(\rho_{\mathrm{m}} + \rho_{\mathrm{r}})).\end{aligned}$$

Since $f''(x) = [(\gamma + \delta)/(\gamma + \delta + 1)]x^{-(\gamma+\delta+2)/(\gamma+\delta+1)} > 0$ for $x > 0$, $f(x)$ is a strictly convex function.
Solving $f'(x) = 0$, we know that the minimum point is $y = [(\gamma + \delta)/(\gamma + \delta + 1)]^{\gamma+\delta+1}$. Since $f(y) = (\gamma + 1)\rho_{\mathrm{r}}/(\rho_{\mathrm{m}} + \rho_{\mathrm{r}}) - 1/(1 + 1/(\gamma + \delta))^{y+\delta} > 0$, we have $f = ((\rho_{\mathrm{m}} - \gamma\rho_{\mathrm{r}})/(\rho_{\mathrm{m}} + \rho_{\mathrm{r}})) > 0$, i.e., $\bar{\pi}^* - \pi^*$.
(ii) Proof of $q^* < \bar{q}^*$:

$$\begin{aligned}\bar{q}^* - q^* = {} & [(\delta)^{(\gamma+1)}\beta\gamma^{-\gamma}(\rho_{\mathrm{m}} + \rho_{\mathrm{r}})]^{1/(\gamma+\delta+1)} \\ & \times\{1 - ((\rho_{\mathrm{m}} - \gamma\rho_{\mathrm{r}})/(\rho_{\mathrm{m}} + \rho_{\mathrm{r}})^{1/(\gamma+\delta+1)}\} > 0.\end{aligned}$$

(iii) Proof of $\bar{a}^* > a^*$:

$$\begin{aligned}\bar{a}^* - a^* = {} & [(1/\delta)^{\delta}\beta\gamma^{\delta+1}(\rho_{\mathrm{m}} + \rho_{\mathrm{r}})]^{1/(\gamma+\delta+1)} \\ & \times\{1 - ((\rho_{\mathrm{m}} - \gamma\rho_{\mathrm{r}})/(\rho_{\mathrm{m}} + \rho_{\mathrm{r}})^{1/(\gamma+\delta+1)}\} > 0.\end{aligned}$$

(iv) Proof of $1 > t_{\max} > t_{\min} \geq 0$:
Since $t_{\max} - t_{\min} = (k_1 + k_2)/\bar{a}^* = \Delta\pi/\bar{a}^* > 0$, we have $t_{\max} > t_{\min}$.
Now let us show $t_{\max} < 1$.
Let

$$\Phi(x) = (\delta + \gamma)x^{1/(\delta+\gamma+1)} + (\rho_{\mathrm{m}}/(\rho_{\mathrm{m}} - \gamma\rho_{\mathrm{r}}))x^{-(\delta+\gamma)/(\delta+\gamma+1)} - (\delta + \gamma + 1)$$

Since

$$\Phi(x) = ((\delta + \gamma)/(\delta + \gamma + 1))x^{-(2\delta+2\gamma+1)/(\delta+\gamma+1)}\{x - (\rho_{\mathrm{m}}/(\rho_{\mathrm{m}} - \gamma\rho_{\mathrm{r}}))\} > 0$$
$$\text{when } x > \rho_{\mathrm{m}}/\left(\rho_{\mathrm{m}} - \gamma\rho_{\mathrm{r}}\right)$$

we have $\Phi((\rho_{\mathrm{m}} + \rho_{\mathrm{r}})/(\rho_{\mathrm{m}} - \gamma\rho_{\mathrm{r}})) > \Phi(\rho_{\mathrm{m}}/(\rho_{\mathrm{m}} - \gamma\rho_{\mathrm{r}}))$.
After rearrangement of above inequality, we have $t_{\max} < 1$.

This page intentionally left blank

DETERMINISTIC HYBRID AND STOCHASTIC COMBINATORIAL OPTIMIZATION TREATMENTS OF AN ELECTRONIC PRODUCT DISASSEMBLY LINE

Seamus M. McGovern and Surendra M. Gupta

ABSTRACT

Disassembly takes place in remanufacturing, recycling, and disposal, with a line being the best choice for automation. The disassembly line balancing problem seeks a sequence that is feasible, minimizes the number of workstations, and ensures similar idle times, as well as other end-of-life specific concerns. Finding the optimal balance is computationally intensive due to exponential growth. Combinatorial optimization methods hold promise for providing solutions to the problem, which is proven here to be NP-hard. Stochastic (genetic algorithm) and deterministic (greedy/hill-climbing hybrid heuristic) methods are presented and compared. Numerical results are obtained using a recent electronic product case study.

Applications of Management Science: In Productivity, Finance, and Operations
Applications of Management Science, Volume 12, 175–197

ISSN: 0276-8976/doi:10.1016/S0276-8976(06)12013-1

INTRODUCTION

Manufacturers are increasingly acting to recycle and remanufacture their post-consumer products due to environmental legislation, public awareness, and extended manufacturer responsibility, as well as the economic attractiveness of reusing products and parts. The first step of product recovery is disassembly. Disassembly is the methodical extraction of valuable parts and materials from discarded products. After disassembly, reusable components are cleaned, refurbished, tested, and directed to inventory for remanufacturing operations. The recyclable materials can be sold to raw-material suppliers, while the residuals are sent to landfills. Obtaining valuable components and materials while minimizing the area required for landfills and reducing the amount of processed toxins introduced into the environment are compelling reasons that disassembly, and specifically the disassembly line, is of such importance and has garnered so much recent interest.

Disassembly has unique characteristics. While possessing similarities to assembly, it is not the reverse of the assembly process (Brennan, Gupta, & Taleb, 1994); therefore, new and efficient approaches and methodologies are needed to effectively perform disassembly line operations. The difficulty in obtaining efficient disassembly line sequence solutions stems from the fact that any solution sequence would consist of a permutation of numbers. This permutation would contain as many elements as there are parts in the product. As such, the observation can be made that the disassembly line balancing problem (DLBP) would appear to be NP-hard and the decision version would appear to be NP-complete (see Tovey, 2000, for a well structured review of complexity including NP-hard and NP-complete). Also of interest, the DLBP is a very recent problem, first formally described in 2002 (Güngör & Gupta, 2002). The importance of the disassembly line's role in end-of-life processing, the contemporary nature of the problem, and its NP-complete characteristics makes the DLBP both academically interesting and scientifically challenging and hence provides much of the primary motivation behind the formulation of this chapter.

This chapter first mathematically defines the DLBP then proves that it is NP-hard, necessitating specialized solution techniques. While exhaustive search consistently provides the problem's optimal solution, its time complexity quickly limits its practicality. Combinatorial optimization is a field that combines techniques from applied mathematics, operations research, and computer science to solve optimization problems over discrete structures. These techniques include: greedy algorithms, integer and linear programming, branch-and-bound, divide and conquer, dynamic programming, local optimization, simulated annealing, genetic algorithms, and approximation algorithms. In this chapter the DLBP is solved using a stochastic metaheuristic (genetic algorithm or GA) and a deterministic hybrid process consisting of a greedy sorting algorithm followed by a hill-climbing heuristic (adjacent element hill climbing (AEHC)).

LITERATURE REVIEW

Many papers have discussed the different aspects of product recovery. Brennan et al. (1994) and Gupta and Taleb (1994) investigated the problems associated with disassembly planning and scheduling. Torres, Gil, Puente, Pomares, and Aracil (2004) reported a study for non-destructive automatic disassembly of personal computers. Güngör and Gupta (1999a, 1999b, 2002) presented the first introduction to the DLBP and developed an algorithm for solving the DLBP in the presence of failures with the goal of assigning tasks to workstations in a way that probabilistically minimizes the cost of defective parts (Güngör & Gupta, 2001). For a review of environmentally conscious manufacturing and product recovery see Güngör and Gupta (1999c). For a comprehensive review of disassembly sequencing see Lambert (2003). McGovern, Gupta, and Kamarthi (2003) first applied combinatorial optimization techniques to the DLBP. McGovern and Gupta applied a novel uninformed general-purpose search heuristic (McGovern & Gupta, 2004) and an ant colony optimization solution (McGovern & Gupta, in press-b) to the DLBP.

THE DLBP MODEL DESCRIPTION

The following notation are used in the chapter:

CT	cycle time; max time at each workstation
d_k	demand; quantity of part k requested
D	demand rating for a solution; also, demand bound for the decision version of DLBP
F	measure of balance for a given solution
h_k	binary value; 1 if part k is hazardous, else 0
H	hazard rating for a solution; also, hazard bound for the decision version of DLBP
I	total idle time for a given solution sequence
j	workstation count $(1, \ldots, NWS)$

k	part identification $(1, \ldots, n)$
n	number of part removal tasks
$\mathbf{N}$	the set of natural numbers $\{0, 1, 2, \ldots\}$
NWS	workstations required for a given solution
P	the set of n part removal tasks
PRT_k	part removal time required for kth task
PS_k	kth part in a solution sequence (i.e., for solution $\langle 3, 1, 2\rangle$ $PS_2 = 1$)
r_k	integer corresponding to removal direction
R	direction rating for a solution; also, direction bound for the decision version of DLBP
R_k	binary value; 0 if part k can be removed in the same direction as part $k+1$, else 1
ST_j	station time; total processing time requirement in workstation j
V	maximum range for a workstation's idle time
$\mathbf{Z}^+$	the set of positive integers; i.e., $\{1, 2, \ldots\}$

In DLBP a solution consists of an ordered sequence of work elements (tasks, components, or parts). For example, if a solution consisted of the n-tuple $\langle 5, 2, 8, 1, 4, 7, 6, 3\rangle$, then part 5 would be removed first, followed by part 2, then part 8, and so on.

While different authors use a variety of definitions for the term "balanced" in reference to assembly (Elsayed & Boucher, 1994) and disassembly lines, we propose the following definition (McGovern & Gupta, 2003; McGovern et al., 2003) that will be used consistently throughout this chapter:

Definition 1. A disassembly line is optimally *balanced* when the fewest possible number of workstations is needed and the variation in idle times between all workstations is minimized. This is mathematically described by

$$\text{Minimize } NWS$$

and then

$$\text{Minimize}[\max(STx) - \min(ST_y)] \forall x, y \in \{1, 2, \ldots, NWS\}$$

Except for some minor variation in the greedy portion of the hybrid approach, both of the combinatorial optimization techniques described here use a similar methodology to address the multi-criteria aspects of DLBP. This evaluation methodology is based on *preemptive* *(lexicographic) goal programming* and is conducted as follows: since measure of balance is the primary consideration in this chapter, additional objectives are only considered subsequently; i.e., the methodologies first seek to select the best performing measure of balance solution; equal balance solutions are then evaluated for hazardous part removal positions; equal balance and hazard measure solutions are evaluated for high-demand part removal positions; and equal balance, hazard measure and high-demand part removal position solutions are evaluated for the number of direction changes. This priority ranking approach was selected over a weighting scheme for its simplicity, ease in re-ranking the priorities, ease in expanding or reducing the number of priorities, due to the fact that other weighting methods can be readily addressed at a later time, and primarily to enable unencumbered *efficacy* (a method's effectiveness in finding good solutions) analysis of the combinatorial optimization methodologies and electronic product problem instance under consideration.

The application investigated in this chapter seeks to fulfill five objectives:

1. minimize the number of workstations and hence, minimize the total idle time,
2. ensure workstation idle times are similar,
3. remove hazardous parts early in the sequence,
4. remove high-demand parts before low-demand parts, and
5. minimize the number of direction changes.

A major constraint is the requirement to provide a feasible disassembly sequence for the product being investigated. Testing a given solution against the precedence constraints fulfills the major constraint of precedence preservation. Minimizing the sum of the workstation idle times also minimizes the total number of workstations. This attains objective 1 and is described by

$$I = \sum_{j=1}^{NWS} (CT - ST_j) \tag{1}$$

Line balancing seeks to achieve *perfect balance* (all idle times equal to zero). When this is not achievable, either Line Efficiency (LE) or the Smoothness Index (SI) is used as a performance evaluation tool (Elsayed & Boucher, 1994).

SI rewards similar idle times at each workstation, but at the expense of allowing for a large (sub-optimal) number of workstations. This is because SI compares workstation elapsed times to the largest ST_j instead of to CT.

(SI is very similar in format to the sample standard deviation from the field of statistics, but using max(STj)|$j \in \{1, 2, \ldots, NWS\}$ rather than the mean of the station times.) LE rewards the minimum number of workstations but allows unlimited variance in idle times between workstations because no comparison is made between ST_js. This chapter makes use of a measure of balance that combines the two and is easier to calculate. The balancing method developed by McGovern and Gupta (2003) and McGovern et al. (2003) seeks to simultaneously minimize the number of workstations while aggressively ensuring that idle times at each workstation are similar, though at the expense of the generation of a nonlinear objective function. The method is computed based on the minimum number of workstations required as well as the sum of the square of the idle times for each of the workstations. This penalizes solutions where, even though the number of workstations may be minimized, one or more have an exorbitant amount of idle time when compared to the other workstations. It provides for leveling the workload between different workstations on the disassembly line. Therefore, a resulting minimum numerical performance value is the more desirable solution, indicating both a minimum number of workstations and similar idle times across all workstations. The measure of balance is represented as

$$F = \sum_{j=1}^{NWS} (CT - ST_j)^2 \tag{2}$$

with the DLBP balancing objective represented as

$$\text{Minimize } Z_2 = \sum_{j=1}^{NWS} (CT - ST_j)^2 \tag{3}$$

Perfect balance is indicated by

$$Z_2 = 0 \tag{4}$$

Note that mathematically, Formula (2) effectively makes Formula (1) redundant due to the fact that it concurrently minimizes the number of workstations.

Theorem 1. Let PRT_k be the part removal time for the kth of n parts where CT is the maximum amount of time available to complete all tasks assigned to each workstation. Then for the most efficient distribution of tasks, the minimum number of workstations, NWS^* satisfies

$$NWS^* \geq \left\lceil \frac{\sum_{k=1}^{n} PRT_k}{CT} \right\rceil \tag{5}$$

Proof. If the above inequality is not satisfied, then there must be at least one workstation completing tasks requiring more than CT of time, which is a contradiction.

Subsequent bounds are shown to be true in a similar fashion and are presented throughout the chapter without proof.

The upper bound for the number of workstations is given by

$$NWS_{\text{nom}} = n \tag{6}$$

Therefore,

$$\left\lceil \frac{\sum_{k=1}^{n} PRT_k}{CT} \right\rceil \leq NWS \leq n \tag{7}$$

The lower bound on F is given by

$$F^* \geq \left(\frac{I}{NWS^*} \right)^2 NWS^* \tag{8}$$

while the upper bound is described by

$$F_{\text{nom}} = \sum_{k=1}^{n} (CT - PRT_k)^2 \tag{9}$$

therefore,

$$\left(\frac{I}{NWS^*} \right)^2 NWS^* \leq F \leq \sum_{k=1}^{n} (CT - PRT_k)^2 \tag{10}$$

The cycle time is then bounded by

$$\max(PRT_k) \leq CT \leq \sum_{k=1}^{n} PRT_k \qquad CT \in \mathbf{N}, \quad \forall k \in P \tag{11}$$

A hazard measure was developed to quantify each solution sequence's performance, with a lower calculated value being more desirable (McGovern & Gupta, 2005). This measure is based on binary variables that indicate whether a part is considered to contain hazardous material (the binary variable is equal to one if the part is hazardous, else zero) and its position in the sequence. A given solution sequence hazard measure is defined as the sum of hazard binary flags multiplied by their position in the solution sequence, thereby rewarding the removal of hazardous parts early in the part removal sequence. This measure is represented as

$$H = \sum_{k=1}^{n}(kh_{PS_k}),\ h_{PS_k} = \begin{cases} 1, & \text{hazardous} \\ 0, & \text{otherwise} \end{cases} \tag{12}$$

with the DLBP hazardous part objective represented as

$$\text{Minimize } Z_3 = \sum_{k=1}^{n}(kh_{PS_k}) \tag{13}$$

The lower bound on the hazardous part measure is given by

$$H^* = \sum_{p=1}^{|HP|} p \tag{14}$$

where the set of hazardous parts is defined as

$$HP = \{k : \ h_k \neq 0 \forall k \in P\} \tag{15}$$

and its cardinality can be calculated with

$$|HP| = \sum_{k=1}^{n} h_k \tag{16}$$

For example, three hazardous parts would give a best-case value of $1+2+3=6$. The upper bound on the hazardous part measure is given by

$$H_{\text{nom}} = \sum_{p=n-|HP|+1}^{n} p \tag{17}$$

or alternatively

$$H_{\text{nom}} = (n|HP|) - |HP| \tag{18}$$

For example, three hazardous parts in a total of twenty would give an H_{nom} value of $18+19+20=57$ or equivalently, $H_{\text{nom}} = (20 \times 3) - 3 = 60 - 3 = 57$. Formulae (14), (17), and (18) are combined to give

$$\sum_{p=1}^{|HP|} p \leq H \leq \sum_{p=n-|HP|+1}^{n} p = (n|HP|) - |HP| \tag{19}$$

A demand measure was developed to quantify each solution sequence's performance, with a lower calculated value being more desirable

(McGovern & Gupta, 2005). This measure is based on positive integer values that indicate the quantity required of a given part after it is removed – or zero if it is not desired – and its position in the sequence. A solution sequence demand measure is then defined as the sum of the demand value multiplied by that part's position in the sequence, rewarding the removal of high-demand parts early in the part removal sequence. This measure is represented as

$$D = \sum_{k=1}^{n} (k d_{PS_k}) \quad d_{PS_k} \in \mathbf{N} \forall_{PS_k} \tag{20}$$

with the DLBP demand part objective represented as

$$\text{Minimize } Z_4 = \sum_{k=1}^{n} (k d_{PS_k}) \tag{21}$$

The lower bound on the demand measure (D^*) is given by Formula (20), where

$$d_{PS_1} \geq d_{PS_2} \geq \cdots \geq d_{PS_n} \tag{22}$$

For example, three parts with demands of 4, 5, and 6, respectively, would give a best-case value of $(1 \times 6) + (2 \times 5) + (3 \times 4) = 28$. The upper bound on the demand measure (D_{nom}) is given by Formula (20), where

$$d_{PS_1} \leq d_{PS_2} \leq \cdots \leq d_{PS_n} \tag{23}$$

For example, three parts with demands of 4, 5, and 6, respectively, would give a worst-case value of $(1 \times 4) + (2 \times 5) + (3 \times 6) = 32$.

Finally, a direction measure was developed to quantify each solution sequence's performance, with a lower calculated value indicating minimal direction changes and a more desirable solution (McGovern & Gupta, in press-b). This measure is based on a count of the direction changes. Integer values represent each possible direction (typically $r = \{+x, -x, +y, -y, +z, -z\}$; in this case$|r| = 6$). These directions are expressed as

$$r_{PS_k} = \begin{cases} +1, & \text{direction} \quad +x \\ -1, & \text{direction} \quad -x \\ +2, & \text{direction} \quad +y \\ -2, & \text{direction} \quad -y \\ +3, & \text{direction} \quad +z \\ -3, & \text{direction} \quad -z \end{cases} \tag{24}$$

and are easily expanded to other or different directions in a similar manner.

The direction measure is represented as

$$R = \sum_{k=1}^{n-1} R_k \quad R_k = \begin{cases} 1, & r_{PS_k} \neq r_{PS_{k+1}} \\ 0, & \text{otherwise} \end{cases} \tag{25}$$

with the DLBP direction part objective represented as

$$\text{Minimize } Z_5 = \sum_{k=1}^{n-1} R_k \tag{26}$$

The lower bound on the direction measure is given by

$$R^* = |r| - 1 \tag{27}$$

For example, for a given product containing six parts that are installed/removed in directions ($-y$, $+x$, $-y$, $-y$, $+x$, $+x$) the resulting best-case value would be 2–1 = 1 (e.g., one possible R^* solution containing the optimal, single-change of product direction would be: $\langle -y, -y, -y, +x, +x, +x \rangle$). In the specific case where the number of unique direction changes is one less than the total number of parts n, the upper bound on the direction measure would be given by

$$R_{\text{nom}} = |r|, \quad \text{where } |r| = n - 1 \tag{28}$$

Otherwise, the measure varies depending on the number of parts having a given removal direction and the total number of removal directions. It is bounded by

$$|r| \leq R_{\text{nom}} \leq n - 1, \quad \text{where}|r| < n - 1 \tag{29}$$

For example, six parts installed/removed in directions ($+ x$, $+x$, $+x$, $-y$, $+x$, $+x$) would give an R_{nom} value of 2 as given by the lower bound of Formula (29) with a solution sequence of, for instance $\langle +x, +x, -y, +x, +x, +x \rangle$. Six parts installed/removed in directions ($-y$, $+x$, $-y$, $-y$, $+x$, $+x$) would give an R_{nom} value of 6–1 = 5 as given by the upper bound of Formula (29) with a solution sequence, for example, of $\langle -y, +x, -y, +x, -y, +x \rangle$. In the special case where each part has a unique removal direction, the measures for R^* and R_{nom} are equal and given by

$$R^* = R_{\text{nom}} = n - 1, \quad \text{where } |r| = n \tag{30}$$

The overall goal is to minimize each of these measures, with a priority of F, then H, then D, and then R. Problem assumptions include the following:

- part removal times are deterministic, constant, and integer (or able to be converted to integers),

- each product undergoes complete disassembly,
- all products contain all parts with no additions, deletions, or modifications,
- each task is assigned to exactly one workstation,
- the sum of the part removal times of all parts assigned to a workstation must not exceed *CT*, and
- part precedence relationships must be enforced.

NP-HARDNESS PROOF

NP-hard can be interpreted as meaning "as hard as the hardest problem in NP." The decision version of DLBP has been shown to be NP-complete (McGovern & Gupta, in press-a) and is described as follows.

Instance. A finite set P of tasks, partial order $<\cdot$ on P, task time $PRT_k \in \mathbf{Z}^+$, hazardous part binary value $h_k \in \{0, 1\}$, part demand $d_k \in \mathbf{N}$, and part removal direction $r_k \in \mathbf{Z}$ for each $k \in P$, workstation capacity $CT \in \mathbf{Z}^+$, number $NWS \in \mathbf{Z}^+$ of workstations, difference between largest and smallest idle times $V \in \mathbf{N}$, hazard measure $H \in \mathbf{N}$, demand measure $D \in \mathbf{N}$, and direction change measure $R \in \mathbf{N}$.

Question. Is there a partition of P into disjoint sets P_A, P_B, …, P_{NWS} such that the sum of the sizes of the tasks in each P_X is CT or less, the difference between largest and smallest idle times is V or less, the sum of the hazardous part binary values multiplied by their sequence position is H or less, the sum of the demanded part values multiplied by their sequence position is D or less, the sum of the number of part removal direction changes is R or less, and it obeys the precedence constraints?

An optimization problem can be proven to be NP-hard in two steps (Garey & Johnson, 1979) by showing that: (1) the NP-complete decision problem is no harder than its optimization problem, and (2) the decision problem is NP-complete.

Theorem 2. DLBP is NP-hard.

Proof. (1) The decision version of DLBP is no harder than its optimization version. Both versions require preservation of the precedence constraints. The optimization version of DLBP asks for a sequence that has the minimum difference between largest and smallest idle times among all disjoint sets, the minimum sum of hazardous part binary values multiplied by their sequence position, the minimum sum of demanded part values multiplied by their sequence position, and the minimum number of part removal direction changes. The decision version includes numerical bounds *V*, *H*, *D*, and *R* as additional parameters and asks whether there exists *NWS* disjoint sets with a difference between largest and smallest idle times no more than *V*, a sum of hazardous part binary values multiplied by their sequence position no more than *H*, a sum of demanded part values multiplied by their sequence position no more than *D*, and a number of part removal direction changes no more than *R*. So long as the difference between largest and smallest idle times, the sum of hazardous part binary values multiplied by their sequence position, the sum of demanded part values multiplied by their sequence position, and the number of part removal direction changes is relatively easy to evaluate, the decision problem can be no harder than the corresponding optimization problem. If we could find a minimum difference between the largest and smallest idle times, a minimum sum of hazardous part binary values multiplied by their sequence position, a minimum sum of demanded part values multiplied by their sequence position, and a minimum number of part removal direction changes for the DLBP optimization problem in polynomial time, then we could also solve the associated decision problem in polynomial time. All we need do is find the minimum idle time and the maximum idle time from all *NWS* subsets, compute their difference, and compare that difference between largest and smallest idle times to the given bound *V*; sum all the hazardous part binary values multiplied by their sequence position and compare that sum to the given bound *H*; sum all the demanded part values multiplied by their sequence position and compare that sum to the given bound *D*; and sum all of the part removal direction changes and compare that sum to the given bound *R*. Therefore, the decision version of DLBP is no harder than its optimization version. (2) From McGovern and Gupta (in press-a), the decision version of DLBP is NP-complete.

Therefore, from (1) and (2), DLBP is NP-hard.

THE GREEDY AND AEHC HEURISTICS

Originally referred to by Garey and Johnson (1979) as *nearest neighbor* (NN) and attributed to a 1965 chapter by Gavett, this search is commonly known as a greedy algorithm (possibly due to a 1985 book chapter reference by Johnson and Papadimitriou where they describe the process as "being 'greedy'"). A greedy strategy always makes the choice that looks the best at the moment. That is, it makes a locally optimal choice in the hope that this choice will lead to a globally optimal solution. Greedy algorithms do not always yield optimal solutions but for many problems, they do. The DLBP Greedy algorithm was built around *first-fit-decreasing* (FFD) rules. FFD rules require looking at each element in a list, from largest to smallest (*PRT* in the DLBP) and putting that element into the first workstation in which it fits without violating precedence constraints. When all of the work elements have been assigned to a workstation, the process is complete. The greedy FFD algorithm in this chapter is further modified with priority rules to meet multiple objectives. During the sorting process, the hazardous parts are prioritized, greedy ranked large removal time to small. The remaining non-hazardous parts are greedy ranked next, large removal times to small. In addition, selecting the part with the larger demand ahead of those with lesser demands breaks any ties for parts with equal part removal times and selecting the part with an equivalent part removal direction breaks any ties for parts also having equal part removal directions. This is done to prevent damage to these more desirable parts. The DLBP Greedy algorithm is able to provide an optimal or near-optimal (minimum) number of workstations. As may be expected with other processes, the more constraints the more likely the optimal solution is found; i.e., the level of performance will generally improve with the number of precedence constraints.

The specific details for this implementation are as follows. The DLBP Greedy algorithm first sorts the list of parts. The sorting is based on part removal times, whether or not the part contains hazardous materials, the subsequent demand for the removed part, and the part removal direction. Hazardous parts are put at the front of the list for selection into the solution sequence. The hazardous parts are ranked from largest to smallest part removal times. The same is then done for the non-hazardous parts. Any ties (i.e., two parts with equal hazard typing and equal part removal times) are not randomly broken, but rather ordered based on the demand for the part, with the higher demand part being placed earlier on the list. Any of these parts also having equal demands is then selected based on their part removal direction being the same as the previous part on the list (i.e., two parts compared during the sorting that only differ in part removal directions are swapped if they are removed in different directions – the hope being that subsequent parts and later sorts can better place parts having equal part removal directions).

Once the parts are sorted in this multi-criteria manner, the parts are placed in workstations in FFD greedy order while preserving precedence.

Each part in the sorted list is examined from first to last. If the part had not previously been put into the solution sequence (as described by ISS_k tabu[1] list data structure), the part is put into the current workstation if idle time remains to accommodate it and as long as putting it into the sequence at that position will not violate any of its precedence constraints. ISS_k is defined as

$$ISS_k = \begin{cases} 1, & \text{assigned} \\ 0, & \text{assigned} \end{cases} \quad \forall k \in P \tag{31}$$

If no workstation can accommodate it at the given time in the search due to precedence constraints, the part is maintained on the sorted list (i.e., its ISS_k value remains 0) and the next part (not yet selected) on the sorted list is considered. If all parts have been examined for insertion into the current workstation on the greedy solution list, a new workstation is created and the process is repeated. The DLBP Greedy heuristic process is run once to generate its solution.

While being very fast and generally very efficient, the FFD-based greedy algorithm is not always able to optimally minimize the number of workstations. In addition, there is no capability to equalize the station times in the workstations; in fact the FFD structure lends itself to filling the earlier workstations as much as possible, often to capacity, while later workstations end up with progressively greater and greater idle times. This results in an extremely poor balance measure.

A two-phase approach to DLBP was developed to quickly provide a near-optimal and feasible balance sequence using a hill-climbing local search heuristic, AEHC (McGovern & Gupta, 2003). AEHC was tailored to DLBP (and is applicable to any bin-packing-type problem, especially those having precedence constraints) to take advantage of knowledge about the problem's format and constraints in order to provide a solution that is better balanced than DLBP Greedy alone but significantly faster than other methodologies. Hill climbing is an iterated improvement algorithm, basically a gradient descent/ascent. It makes use of an iterative greedy strategy, which is to move in the direction of increasing value. A hill-climbing algorithm evaluates the successor states and keeps only the best one. AEHC is designed to consider swapping each task in every workstation only with each task in the next adjacent workstation in search of improved balance. It does this while preserving precedence and not exceeding CT in any workstation. Only adjacent workstations are compared to enable a rapid search and since it is deemed unlikely that parts several workstations apart can be swapped and still preserve the precedence of all of the tasks in-between. As shown in the Fig. 1 example, a part has a limited number of other parts it can be considered with for exchange. Using AEHC, part 6 would be considered for exchange only by parts 4 and 5, and would consider exchanges only with parts 8–10.

The neighborhood definition and search details of AEHC are as follows. After the DLBP greedy algorithm generates a minimum-*NWS* solution that is feasible, the AEHC heuristic is applied to improve the balance. AEHC does this by going through each task element (part) in each workstation and comparing it to each task element in the next adjacent workstation. If the two task elements can be exchanged while preserving precedence, without exceeding either workstations available idle time, and with a resulting improvement in the overall balance, the exchange is made and the resulting solution sequence is saved as the new solution sequence. This process is repeated until task elements of the last workstation have been examined.

Hill climbing is typically continuously run on subsequent solutions for as long as is deemed appropriate or acceptable by the user or until it is no longer possible to improve, at which point it is assumed that the local optima has been reached. Repeating the AEHC method in this way provides improved balance with each iteration. During its development AEHC was tested both ways; run only once after the greedy solution was generated, as well as run until the local optima was obtained (a single AEHC iteration has several benefits including the observation that AEHC was seen to typically provide its largest single balance performance improvement in the first iteration). Additionally, a single iteration of AEHC may be recommended for the real-time solution of a very large DLBP on a dynamic mixed-model and mixed-product disassembly line. In this chapter, AEHC was continuously run on subsequent solutions until no improvements could be seen.

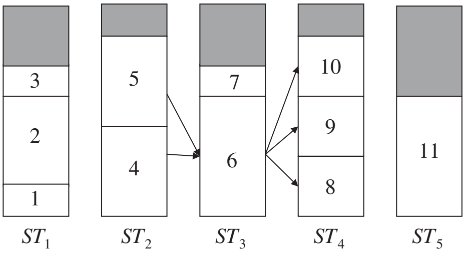


Fig. 1. AEHC Example.

THE DLBP GENETIC ALGORITHM

A GA (a parallel neighborhood, stochastic-directed search technique) provides an environment where solutions continuously crossbreed, mutate, and compete with each other until they evolve into an optimal or near-optimal solution. Due to its structure and search method, a GA is often able to find a global solution, unlike many other heuristics that use hill climbing to find a best solution nearby resulting only in a local optima. In addition, a GA does not need specific details about a problem nor is the problem's structure relevant; a function can be linear, nonlinear, stochastic, combinatorial, noisy, etc.

GA has a solution structure defined as a *chromosome*, which is made up of *genes* and generated by two *parent* chromosomes from the *pool* of solutions, each having its own measure of *fitness*. New solutions are generated from old using the techniques of *crossover* (sever parents genes and swap severed sections) R_x and *mutation* (randomly vary genes within a chromosome) R_m. Typically, the main challenge with any GA implementation is determining a chromosome representation that remains valid after each generation.

For DLBP, the chromosome (solution) consisted of a sequence of genes (parts). A pool, or *population*, of size N was used. Only feasible disassembly sequences were allowed as members of the population or as offspring. The fitness was computed for each chromosome using the preemptive (lexicographic) goal programming method for solution performance determination as described previously.

DLBP-Specific GA Architecture

The GA for DLBP was constructed as follows. An initial, feasible population was taken from a hot-start list and the fitness of each chromosome in this generation was calculated (the primary technique of randomly generating the initial population was ineffective due to the electronic product instance's relatively large size and numerous precedence constraints, so four feasible and diverse solutions were manually generated to hot start DLBP GA). An even integer of $R_x \cdot N$ parents was randomly selected for crossover to produce $R_x \cdot N$ offspring (offspring make-up $(R_x \cdot N \cdot 100)$ percent of each generation's population). An elegant crossover, the precedence preservative crossover (PPX) developed by Bierwirth, Mattfeld, and Kopfer (1996) was used to create the offspring. PPX first creates a mask (one for each child, every generation). The mask consists of random 1 and 2s indicating which parent part information should be taken from. If, for example, the mask for child 1 reads 221211, the first two parts (i.e., from left to right) in parent 2 would make up the first two genes of child 1 (and these parts would be stricken from the parts available to take from both parent 1 and 2); the first available (i.e., not stricken) part in parent 1 would make up gene three of child 1; the next available part in parent 2 would make up gene four of child 1; the last two parts in parent 1 would make up genes five and six of child 1. This technique is repeated using a new mask for child 2.

After crossover, mutation is randomly conducted. Mutation was occasionally (based on the R_m value) performed by randomly selecting a single child then exchanging two of its disassembly tasks while ensuring precedence is preserved. The $R_x \cdot N$ least fit parents are removed by sorting the entire parent population from worst-to-best based on fitness.

Since the GA saves the best parents from generation to generation and it is possible for duplicates of a solution to be formed using PPX, the solution set could contain multiple copies of the same answer resulting in the algorithm potentially becoming trapped in a local optima. This becomes more likely in a GA with solution constraints (such as precedence requirements) and small populations, both of which are seen in the study in this chapter. To avoid this, DLBP GA was modified to treat duplicate solutions as if they had the worst fitness performance (highest numerical value), relegating them to replacement in the next generation. With this new ordering, the best unique $(1–R_x)N$ parents were kept along with all of the $R_x \cdot N$ offspring to make up the next generation then the process was repeated. To again avoid becoming trapped in local optima, DLBP GA – as is the case with many combinatorial optimization techniques – was run not until a desired level of performance was reached but rather for many generations (as determined by the user). Since DLBP GA always keeps the best solutions thus far from generation to generation, there is no risk of solution drift or bias and the possibility of mutation allows for a diverse range of possible solution space visits over time.

DLBP-Specific GA Qualitative Modifications

DLBP GA was modified from a general GA in several ways. Instead of the worst portion of the population being selected for crossover, in DLBP GA all of the population was (randomly) considered for crossover. This better enables the selection of nearby solutions (i.e., solutions similar to the best solutions to-date) common in many scheduling problems. Also, mutation was performed only on the children, not the worst parents. This was done to address the small population used in DLBP GA and to counter PPX's tendency to duplicate parents. Finally, duplicate children are sorted to make their deletion from the population likely since there is a tendency for the creation of duplicate solutions (due to PPX) and due to the small population saved from generation to generation.

DLBP-Specific GA Quantitative Modifications

A small population was used (20 versus the more typical 10,000–100,000) to minimize data storage requirements and simplify analysis while a large number of generations were used (10,000 versus the more typical 10–1,000) to compensate for this small population while not being so large as to take an excessive amount of processing time. Lower than the recommended 90% (Koza, 1992), a 60% crossover was selected based on test and analysis. Developmental testing indicated that a 60% crossover provided better solutions and did so with one-third less processing time. Previous assembly line balancing literature that indicated best results have typically been found with crossover rates of from 0.5 to 0.7 also substantiated the selection of this lower crossover rate. A mutation was performed about 1% of the time. Although some texts recommend 0.01% mutation while applications in journal chapters have used as much as 100% mutation, it was found that 1.0% gave excellent algorithm performance for the DLBP.

ELECTRONIC PRODUCT CASE STUDY

Both combinatorial optimization techniques were used to provide a solution to the DLBP instances based on the disassembly sequencing problem presented by Gupta, Evren, and McGovern (2004). The growth of cellular telephone use and rapid changes in technology and features has prompted the entry of new models on a regular basis while, according to Collective Good International, one hundred million cellular telephones are discarded each year. Unwanted cell phones typically end up in landfills and usually contain numerous hazardous parts that may contain mercury, cadmium, arsenic, zinc, nickel, lead, gallium arsenide, and beryllium, any of which can pose a threat to the environment. Gupta et al. (2004) selected a 2001 model year Samsung SCH-3500 cell phone for disassembly analysis. The result is an appropriate, real-world instance consisting of $n = 25$ components having several precedence relationships. The data set includes a paced disassembly line operating at a speed which allows $CT = 18$ s per workstation. Collected data on the SCH-3500 is listed in Table 1. Demand is estimated based on part

Table 1. SCH-3500 Cellular Telephone Parts and their Properties.

Task	Part Removal Description	Time	Hazardous	Demand	Direction
1	Antenna	3	Yes	4	$+y$
2	Battery	2	Yes	7	$-y$
3	Antenna guide	3	No	1	$-z$
4	Bolt (type 1) a	10	No	1	$-z$
5	Bolt (type 1) b	10	No	1	$-z$
6	Bolt (type 2) 1	15	No	1	$-z$
7	Bolt (type 2) 2	15	No	1	$-z$
8	Bolt (type 2) 3	15	No	1	$-z$
9	Bolt (type 2) 4	15	No	1	$-z$
10	Clip	2	No	2	$+z$
11	Rubber seal	2	No	1	$+z$
12	Speaker	2	Yes	4	$+z$
13	White cable	2	No	1	$-z$
14	Red/blue cable	2	No	1	$+y$
15	Orange cable	2	No	1	$+x$
16	Metal top	2	No	1	$+y$
17	Front cover	2	No	2	$+z$
18	Back cover	3	No	2	$-z$
19	Circuit board	18	Yes	8	$-z$
20	Plastic screen	5	No	1	$+z$
21	Keyboard	1	No	4	$+z$
22	LCD	5	No	6	$+z$
23	Sub-keyboard	15	Yes	7	$+z$
24	Internal IC	2	No	1	$+z$
25	Microphone	2	Yes	4	$+z$

value and/or recycling value; part removal times and precedence relationships (Fig. 2) were determined experimentally. Part removal times were repeatedly collected until a consistent part removal performance was attained.

The resulting search space is 25! or 1.55112×10^{25}. Note that in 2005, companies that refurbished cellular telephones typically paid consumers \$2–20 per phone for the more popular Motorola and Nokia cellular telephones while cellular telephone recyclers (that may only extract precious metals such as gold from the circuit boards) paid between \$1 and 6 per phone.

NUMERICAL RESULTS

DLBP GA and the DLBP Greedy/AEHC hybrid were both written in C++ and run on a 1.6 GHz PM x86-family workstation. The developed algorithms were investigated on a variety of test cases for verification and

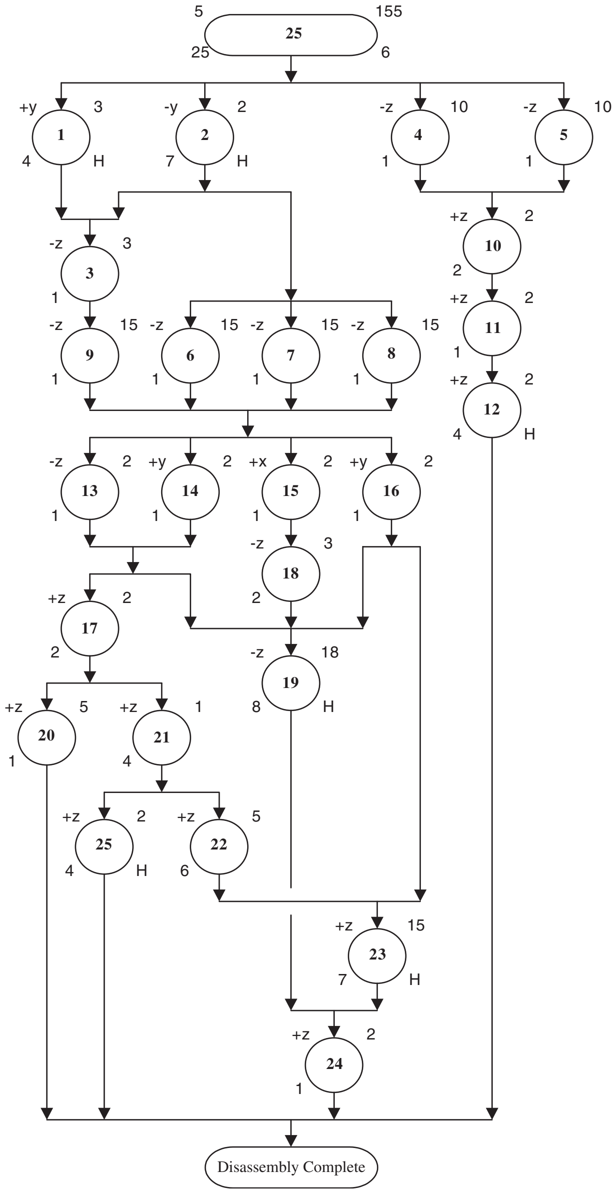


Fig. 2. Cellular Telephone Precedence Relationships.

Table 2. Comparison of Averaged Greedy/AEHC and GA Solutions for the Electronic Product Example.

Solution	Time	*NWS*	*F*	*H*	*D*	*R*
Hybrid	0.007	10	199	89	973	10
GA	0.863	10.0	81.0	78.2	915.2	10.2

validation purposes. They were then compared using the case study from the literature.

For the results in Table 2, the DLBP Greedy/AEHC heuristic hybrid was run three times to obtain an average computation time, while DLBP GA was run five times (unlike Greedy/AEHC, GA is highly probabilistic, so additional runs were conducted to achieve an average measure of performance along with best and worst case). DLBP GA typically found feasible solutions to the electronic product data set in less than 1 s. DLBP Greedy/AEHC found its solution extremely quickly, regularly taking less than 1% of the time of DLBP GA. For comparison, a much smaller problem ($n = 10$) solved by exhaustive search on the same workstation took almost 10 s. At this rate, solving the electronic product problem to optimality using exhaustive search would take over 250 million years. Neither technique was able to generate the optimal number of workstations or the optimal balance (with the assumption that $NWS = 9$ and $F = 9$ is optimal; Gupta et al. (2004)). Using the four manually generated solutions for a hot start, DLBP GA gave five different answers over five runs, seeming to indicate that the hot start solutions were adequate in number and diversity. DLBP GA, on average, provided a better measure of balance and slightly better *H* and *D* placement, while DLBP Greedy/AEHC gave slightly better *R* placement. For a more in-depth study of these methods and analysis on more diverse problem sets (see McGovern & Gupta, in press-a and McGovern & Gupta, 2005). Although these results are not optimal, this is more a reflection of the challenges posed even by seemingly simple disassembly problems more than an indication of any limitations of these techniques. Note that the inclusion of additional precedence constraints will increasingly move both methods toward optimal.

CONCLUSIONS

With efficient disassembly being an essential component of cost-effective reverse logistics and a necessary element of a productive and ecologically aware end-of-life supply chain, two very fast combinatorial optimization approaches to the multiple-criteria DLBP were developed and compared. Though distinct in their approaches, both methods provide feasible solutions to disassembly line problem instances, with one using a stochastic distributed search and the other a hybrid pairing of two deterministic searches. These diverse solution-generating techniques appear well suited to the multiple-criteria decision-making problem format as well as to the solution of problems with nonlinear objectives. In addition, they are ideally suited to integer problems that generally do not lend themselves to application of traditional mathematical programming techniques.

NOTES

1. Although the name recalls tabu search, the ISS_k tabu list is similar to that used in the Ant Colony Optimization metaheuristic including, for example, the absence of any *aspiration* function.

REFERENCES

Bierwirth, C., Mattfeld, D. C., & Kopfer, H. (1996). On permutation representations for scheduling problems, parallel problem solving from nature. In: H. M. Voigt, W. Ebeling, I. Rechenberg, & H. P. Schwefel (Eds), *Lecture notes in computer science* (Vol. 1141, pp. 3.10–3.18). Berlin: Springer.

Brennan, L., Gupta, S. M., & Taleb, K. N. (1994). Operations planning issues in an assembly/disassembly environment. *International Journal of Operations and Production Planning*, *14*(9), 57–67.

Elsayed, E. A., & Boucher, T. O. (1994). *Analysis and control of production systems.* Upper Saddle River, NJ: Prentice-Hall.

Garey, M. R., & Johnson, D. S. (1979). *Computers and intractability: A guide to the theory of NP-completeness.* New York: W. H. Freeman and Company.

Güngör, A., & Gupta, S. M. (1999a). A systematic solution approach to the disassembly line balancing problem. *Proceedings of the 25th International Conference on Computers and Industrial Engineering*, New Orleans, LA, March, pp. 70–73.

Güngör, A., & Gupta, S. M. (1999b). Disassembly line balancing. *Proceedings of the 1999 Annual Meeting of the Northeast Decision Sciences Institute*, Newport, RI, March, pp. 193–195.

Güngör, A., & Gupta, S. M. (1999c). Issues in environmentally conscious manufacturing and product recovery: A survey. *Computers and Industrial Engineering*, *36*(4), 811–853.

Güngör, A., & Gupta, S. M. (2001). A solution approach to the disassembly line problem in the presence of task failures. *International Journal of Production Research*, *39*(7), 1427–1467.

Güngör, A., & Gupta, S. M. (2002). Disassembly line in product recovery. *International Journal of Production Research*, *40*(11), 2569–2589.

Gupta, S. M., Evren, E., & McGovern, S. M. (2004). Disassembly sequencing problem: A case study of a cell phone. *Proceedings of the 2004 SPIE International Conference on Environmentally Conscious Manufacturing IV*, Philadelphia, PA, October, pp. 43–52.

Gupta, S. M., & Taleb, K. N. (1994). Scheduling disassembly. *International Journal of Production Research*, *32*, 1857–1866.

Koza, J. R. (1992). *Genetic programming: On the programming of computers by the means of natural selection*. Cambridge, MA: MIT Press.

Lambert, A. J. D. (2003). Disassembly sequencing: A survey. *International Journal of Production Research*, *41*(16), 3721–3759.

McGovern, S. M., & Gupta, S. M. (2003). Greedy algorithm for disassembly line scheduling. *Proceedings of the 2003 IEEE International Conference on Systems, Man, and Cybernetics*, Washington, DC, October, pp. 1737–1744.

McGovern, S. M., & Gupta, S. M. (2004). Demanufacturing strategy based upon metaheuristics. *Proceedings of the 2004 IIE Industrial Engineering Research Conference*, Houston, TX, May, CD-ROM.

McGovern, S. M., & Gupta, S. M. (2005). Local search heuristics and greedy algorithm for balancing the disassembly line. *The International Journal of Operations and Quantitative Management*, *11*(2), 91–114.

McGovern, S. M., & Gupta, S. M. (in press-a). A balancing method and genetic algorithm for disassembly line balancing. *European Journal of Operational Research*.

McGovern, S. M., & Gupta, S. M. (in press-b). Ant colony optimization for disassembly sequencing with multiple objectives. *The International Journal of Advanced Manufacturing Technology*.

McGovern, S. M., Gupta, & S. M., Kamarthi, S. V. (2003). Solving disassembly sequence planning problems using combinatorial optimization. *Proceedings of the 2003 Annual Meeting of the Northeast Decision Sciences Institute*, Providence, RI, March, pp. 178–180.

Torres, F., Gil, P., Puente, S. T., Pomares, J., & Aracil, R. (2004). Automatic PC disassembly for component recovery. *International Journal of Advanced Manufacturing Technology*, *23*(1–2), 39–46.

Tovey, C. A. (2000). Tutorial on computational complexity. *Interfaces*, *32*(3), 30–61.

This page intentionally left blank

USING "MIRROR WORLDS" TO SUPPORT SUPPLY NETWORK MANAGEMENT

Daniel O'Leary

ABSTRACT

"Mirror Worlds" were suggested by David Gelernter based on a bold assertion: "You will look into a computer screen and see reality." With mirror worlds, managers could be proactive, anticipating what might happen and acting accordingly, instead of waiting till events happen and then reacting. This paper extends the notion of mirrors worlds to supply chain management. In the case of supply chain management, managers could test the impact of making changes in their supply chains to study the impact.

However, mirror worlds could be extended to help actually monitor and manage supply chains to respond and adapt to changes in the world that affected the supply chain. In particular, mirror worlds could be "real" worlds if control for some of the activities between supply chain participants is in effect "turned over" to the mirror world. In that case, the mirror world would show the actual world, with the system making many of the decisions.

1. INTRODUCTION

"Mirror Worlds" were suggested by David Gelernter based on a bold assertion: "You will look into a computer screen and see reality." With mirror

Applications of Management Science: In Productivity, Finance, and Operations
Applications of Management Science, Volume 12, 199–212

ISSN: 0276-8976/doi:10.1016/S0276-8976(06)12014-3

worlds, managers could be proactive, anticipating what might happen and acting accordingly, instead of waiting till events happen and then reacting.

However, Gelernter's view of how mirror worlds could be used in a supply network environment was not elicited in his book. In addition, there have been few contributions to the academic literature focusing on mirror supply network worlds. As a result, one purpose of this paper is to summarize one such vision integrating supply networks and mirror worlds.

In addition, the notion of "mirror worlds" is a bit passive. Rather than just mirroring the "real world," portions of the mirror world might become the real world. For example, rather than just building visibility into a simulation of what could happen in an organization, parts of the mirror world and real world could be merged. If the mirror world is worth its salt as a mirror world then at least parts of it could become parts of the real world and not just mirrors. For some sets of actions, the mirror world would take over decision-making and become the real world.

Is There an Interest in Mirror Worlds for the Supply Chain?

Gelernter's book did not build a mirror world for supply chains. One reason for not building such a model could have been interest. So, is there really an interest in such a view? Recently, Claus Heinrich of the large enterprise resource planning system software company, SAP, noted (SAP, 2001)

> The ultimate goal is to create a truly adaptive supply network that can sense and respond to rapidly evolving conditions so that partners can intelligently cooperate to keep demand and supply in close alignment and efficiently coordinate the fulfillment process. We believe that intelligent agents will be the key to resolving the increasing challenges companies are faced with in participating and managing global adaptive supply networks. (Thomas, 2002).

Accordingly, there appears to be substantial corporate interest in such a model. But, what needs to be embedded in the model?

Dynamic Supply Networks

Teams of companies have announced the development of software designed to attack the problem of supply chain networks. For example,

> … SAP today announced the enhancement of mySAPSupply^{TM} Chain Management (mySAP SCM) to manage adaptive supply chain networks through use of new intelligent agent technology. Adaptive supply networks are an evolution of supply chains and will uniquely combine global visibility, event management, adaptive planning and execution,

> and dynamic collaboration. An adaptive supply network will provide business partner integration and dynamic collaboration through portals and exchanges. http://www11.sap.com/company/press/press.epx?pressID = 179

This statement also suggests that the technology will not just be designed to mirror the world, but actually be used to support and make critical decisions.

Purpose of This Paper

The purpose of this paper is three-fold. First, this paper applies the concept of mirror worlds to supply chain management, in order to elicit a mirror world vision of supply chain management. Second, it suggests actually using the mirror world as part of the real world to help better monitor and manage supply chains. Mirror world capabilities can be used in the real world to execute activities in real time. Third, this paper investigates the extent to which mirror worlds or what appear to be mirror worlds have been used in practice to investigate and improve supply chains, by examining a case study about Procter & Gamble (P&G) and SAP.

This Paper

This paper proceeds in the following way. Section 2 provides some background information on mirror worlds. Section 3 investigates the notion of a supply network, in contrast to a supply chain. Section 4 analyzes some of the different types of events needed throughout a supply network, placing them in multiple categories. Section 5 analyzes the use of intelligent agents, using some examples from practice. Section 6 investigates data sources for mirror worlds, including accounting data, radio frequency identification data (RFID), global satellite positioning data, and Internet data. Section 7 integrates supply network events, intelligent agents, and data sources into the notions of mirror worlds. Section 8 brings it all together with an illustration of the concepts using P&G as an example. Section 9 summarizes the paper.

2. MIRROR WORLDS (GELERNTER, 1992)

Mirror worlds are computer software models of some portion of the world. Mirror worlds gather information from whatever sources they can which can facilitate modeling the real world. Mirror worlds try to model and mimic the way the world works with all of its complex interactions between the different actors. Mirror worlds "mirror" those actors with multiple interacting programs. Mirror worlds are concerned with both the overall view of the interaction and emergent behavior, referred to as "top sight," and the detailed behavior of particular agents, referred to as the "ant's view."

Data are funneled into the mirror world from many sources, since mirror worlds depend heavily on data flow. Data can include classic accounting sources, recently available Internet sources or emerging RFID.

Mirror worlds use computer-based intelligent agents as the actors. Monitoring individual agents provides the ant's view. Agents are designed to follow particular rules of behavior, e.g., "first items in, are the first items out," or "last items in are the last items out." The rules built into the agents in the mirror world often are designed to mimic those in the real world. Alternatively, different rules can be parameterized and studied so that the network can be "optimized." Optimization can occur for either outcomes associated with individual agents or for the entire supply network. Interactions between the agents create the emerged behavior or "top sight," through the development of a network that represents the supply structure.

However, the notion of the mirror world can be extended beyond that of "mimicking" reality. Instead of mimicking, with the right data and the right models, control of the portion of the world under consideration can be delegated to the mirror world. In that setting, the "mirror" world becomes the "real" world.

3. SUPPLY CHAINS OR SUPPLY NETWORKS?

Manufacturing organizations try to manage supplies of raw materials and other manufacturing goods to their firms. Suppliers play different roles, ranging from arms-length third parties to tightly integrated supplier–manufacturer links. That flow of materials from different parties often is referred to as the supply chain.

The importance of the supply chain is being recognized over time. For example, rather than just firm against firm, economic competition increasingly is supply chain against supply chain (DStar, 2001).

However, increasingly, organizations realize that instead of supply chains, organizations constitute supply networks. For example, as noted by P&Gs, director of supply network innovation, "Chain connotes something that is sequential, that requires handing off information in sequence. We believe that it (the supply network) has to operate like a network, like an internet, so everybody has visibility to the information." (Anthes, 2003).

Since chains are sequential, but networks are not necessarily even acyclic, supply networks are combinatorially more complex than supply chains. As a result, supply networks are increasingly seen as an issue for the so-called "complex adaptive systems."

Complex Adaptive Systems

Complex adaptive systems are networks of actors or firms (nodes), each acting on their own behalf. Arcs in the network indicate the movement of goods or information between the nodes or both.

In complex adaptive networks, the networks themselves have emergent behavior, based on the interaction of the behaviors at each of the nodes. This is true even if the amount of intelligence at the nodes is limited. Following even simple rules, systems composed of interacting actors generates behavior for the system, as a consequence of the actor interactions.

Supply networks are complex adaptive networks. The actors are the individual firms and people involved in bringing materials to their customers. Knowing that the supply network is a complex adaptive network means that will have emergent behavior, which we can study to see what rules will generate the kind of outcomes that we are interested in.

Limitations of Supply Network Integration

Unfortunately, sometimes supply networks or portions of supply networks are not integrated. This lack of supply network integration can play a critical role in limiting the flow of information, for a number of those reasons. First, if the systems are not integrated then knowledge of information can be slowed, lost or not possible. Without integration, information cannot flow from one node in the supply network to another. Second, even if systems are integrated, if the same ontology or taxonomy is not used then communication between systems will be slowed, lost or not possible. If one company defines a sale, as on delivery of goods, while another defines a sale on receipt of payment, the sales will not be the same, and information will be confused in the system (e.g., McAfee & McFarland, 2004). Third, without access to multiple different databases could limit the ability of agents, human or computer-based, to make sense out of the impact of information sets. For example, in order to understand the impact of weather would require not only knowledge of the weather, but which, if any, shipments would be affected by the weather. The Internet and the movement toward Internet-based standards has begun to facilitate database integration.

Fourth, the lack of integration may influence the ability of a system to be adaptive. If information flows too slowly then adaptations may be insufficient or inappropriate or too late. If adaptive changes are too late, e.g., after a weather storm is over, then the adaptations may not be affective.

4. SUPPLY NETWORK EVENT TYPES

Supply networks need to be adaptive. In order to be adaptive, they need to be able to monitor "events" within the supply network and respond to those events. "Events" are relevant occurrences or happenings in the supply network. Typically events must be responded to, adapted to or linked to other events. For example, receiving goods is likely to link to paying for those goods. In order to monitor and respond to supply network events requires defining the event types and the range of events that the firm will examine as part of their supply network. There are a broad base of potential event types that might be experienced in a supply network, including accounting, delivery interruption, and other types of events.

Accounting Events

Historically, the primary interest has been in "accounting" type events such as "goods delivered," "goods received," "sales," "purchases," and other similar events. Each of those kinds of events has direct and measurable consequences on the firm. Further, these types of events use information of a particular type, that of monetary-based accounting information. Accounting has a well-defined set of actions and activities.

Delivery Interruption

However, accounting events do not cover the range of events necessary if the system is to be adaptive to changing circumstances in the supply network. In particular, the set of events could be expanded to other types of events, depending on what types of intelligent responses are built into the system. For example, events that relate to interruptions of the supply network could be captured. These might be based on the form of travel delivering the goods, including "ship interruption," "train delivery interruption," and "truck delivery interruption." Within each of those types of interruption there can be additional types of information relating to key characteristics of the impact on the goods. For example, within truck delivery, there could be events ranging from "flat tire – minor," to "truck and goods destroyed." The first of those two events would be reflective of a minor shift in delivery time, while the second, would require replenishing the goods. Different types of events would lead to different sets of actions by various actors in the supply network.

Events could also be expanded to anticipate an impact on particular forms of delivery. For example, weather and traffic information could be monitored. Weather, could be broken down into different types that could impact goods delivery. Accordingly, snow, rain, tornados, and hurricanes could be monitored for the potential impact on goods delivery.

Similarly, traffic in particular settings could be monitored for its impact on particular deliveries. Delays could be categorized according to a number of different taxonomies. Factors affecting the traffic quality include "road construction," "special events," and "natural disaster." Other factors could be related to particular local activities. For example, in Los Angeles, "filming" is an event that can clog the highways and limit access to particular facilities during what would be ordinary delivery times.

Carrier, Trailer, and Loading Events (e.g., Maloney, 2005)

There are a number of carrier, trailer and loading-based events. Carrier-related events include, "carrier accepts assignment" and "carrier does not accept assignment." Trailer-related events include "trailer available," "trailer unavailable," "trailer departure," "trailer arrival," and others. Loading-based events can include "loading begins," "loading completed," and other loading events.

Other Types of Events

Other types of events could also be elicited, based in part of the particular supply network and its needs. For example, there can be unloading interruptions. Unloading interruptions could be a function of "inappropriate equipment," "day of week," "hour of day," "contract agreements," and others. Associated with each type of unloading interruptions could be additional sets of events.

Similarly, in contrast to delivery interruption, there could be "receiving interruptions," depending in part on the perspective being taken. Further, identifying, detecting, and eliminating wasted resources in the supply network could also be set as events.

Importance of Event Definitions

Events definitions provide insight into what will be monitored and managed. If the only definitions are accounting-based, then the capabilities will be limited to those accounting events. However, by expanding the scope to a broader base of events allows the supply network to better manage all its resources. Events can be managed early in their life cycle to help solve minor problems before they turn into major problems. For example, determining if an assignment is not accepted by a carrier early will limit problems of a non delivery later.

5. INTELLIGENT AGENTS

In "real life" agents are those who are authorized to act for others (e.g., Croft, 1997). Computer-based intelligent agents are agents that are computer programs, software that typically has a single function.

Radjou et al. (2002) argue that intelligent agents have multiple properties, including

- Cooperation – agents can cooperate with other agents to work toward a goal.
- Autonomy – agents can work without substantial intervention.
- Reactivity – agents can understand their environment and react accordingly.
- Adaptability – agents can adapt to alternative goals.
- Learning – agents can learn, either as an individual agent or as a system.
- Proactivity – agents have goal directed activity.

Generic Supply Network Roles

Within supply networks there are a number of generic roles for intelligent agents, including the following:

- Sorting agents determine to which agent different events should be funneled to.
- Demand agents determine how much product should be funneled to a particular location. These agents take demand signals and couple it with other events, such as advertising, and predict demand.
- Auction bidding agents, determine what bid is appropriate for some auction and act on executing that auction.

- Interruption analysis agents take in information about shipping interruptions and analyze that interruption to first determine if the interruption will impact the supply network, and if necessary, develop alternative approaches to mitigate the impact of the interruption.

Other roles can be developed based on the particular needs being addressed. For example, if the model has carrier interaction, agents could be concerned with choosing carriers, rescheduling carriers, etc.

Example – Cisco (1998)

One of the most visible adopters of intelligent agents into their customer relationship has been Cisco, who actually purchased a company that made agents to facilitate their own use. Cisco has introduced a number of agents including the following:

- Configuration agents that verify router configurations.
- Pricing agents that help customers view prices.
- Lead Time agents that check expected lead times.
- Status agents that monitor the status of orders.
- Orders Extract agents that take order information at Cisco and download it for local use.

6. MIRROR WORLD DATA SOURCES FOR SUPPLY NETWORKS

Mirror worlds are heavily data driven, using the data to drive agent decision-making where feasible. In a supply network environment, mirror world data sources for supply networks are likely to be from a number of sources, including accounting data, Global Positioning Systems (GPS) and RFID, Internet-based data, and other sources.

Accounting Data

Accounting data can provide substantial input to a supply network mirror world. Supply network data accounting data generated includes

- When, what, and where goods were shipped.
- When, what, and where goods were received.

- When, what, and where goods were recognized as a sale.
- When, what, and where goods were recognized as a purchase.

Throughout, accounting data is focused on economic transactions that occur throughout the supply network. However, that is not to say that other data in the supply network does not have economic consequences, and should not be gathered. Much of this other data can provide the ability of the networks to be adaptive.

GPS and RFID

RFID and GPS have become an increasingly, important source of information about the location of particular materials, or shipments of materials. Information about location can provide improved visibility about shipments or goods. RFID can provide different objects of interest with their own identity. For example, RFID can be used at the individual item level, the case level, the pallet level or even the truck or other form of shipment level. Using RFID and GPS we can facilitate location of those particular objects.

Data generated by GPS and RFID can provide much data to understand what is going on in a supply network, e.g., where are shipments or how long have they been "stuck"? Using GPS and RFID supply network participants can keep track of particular units, knowing where they are located, and keeping track of that location. With that kind of knowledge, goods can become visible to the entire supply network. Such visibility can make the goal of adaptation clear, and even provide data for generation of strategies.

Internet-Based Data

The Internet can provide a number of different kinds of data, such as weather and traffic data. For example, http://www.weather.com provides ongoing information about weather that could be used, while http://traffic-info.lacity.org/ provides traffic information for Los Angeles. Intelligent agents gathering such data can be used to anticipate the flow of goods through the supply chain. In order for agents to fully employ weather and traffic data, they would need to integrate that information with accounting data and GPS and RFID data about supply location. As a result, it would be necessary to integrate the computer-based agents, the events being monitored and the corresponding data sources.

7. INTEGRATING INTELLIGENT AGENTS, EVENT MONITORING, AND DATA SOURCES

Key building blocks in the development of mirror worlds are intelligent computer-based agents, event monitoring, and a keen focus on the appropriate data sources. However, those building blocks do not stand alone, but need to be integrated with each other.

Intelligent Agents and Event Monitoring

In an integrated system, the events would be captured and funneled to intelligent agents that would monitor and sort the events. Intelligent agents would monitor the events, for example, to determine if the events were nonroutine and problematic or routine and non-problematic. Those same intelligent agents could provide responses to routine problems and funnel problematic events to the appropriate source. After events were sorted then they could be sent to the next layer of agents.

Routine problems could be attacked using rules, or sets of rules as part of a rule-based knowledge-based system. In addition, agents could process simulations or other mirror worlds under different scenarios to address different event types. For example, based on the events discussed earlier, different interruption types could be addressed by different agent types. Accordingly, agents could specialize in truck shipping or train shipping. As a result, agents could have knowledge of alternative trucking options or alternative shipping options. Such agents would have knowledge of alternative shipping opportunities. This also would require intelligent agents to access the appropriate data.

Intelligent Agents and Data Sources

Just as events are often tied to particular data sources, intelligent agents must be given access to the appropriate data sources. For example, intelligent agents funneled events related to shipment interruption would need access to data about the interruption. In order to fully respond to a particular type of weather, agents would need access to potential impact of weather on particular locations and the corresponding location of the transport media. As a result, information from satellite-base GPS and RFID could be integrated with information about weather. Further, in order to fully leverage direct links with data sources, agents could be specially designed for particular data sources.

Event Monitoring and Data Sources

Monitoring of specific events (e.g., delivery interruptions in a particular location) can often be traced to monitoring particular data sources for particular outcomes (e.g., traffic web site for slowed traffic levels). Accordingly, there is a close coupling of events and data sources. Agents could be designed to leverage information about accessing particular databases. Event monitoring could leverage xml data exchange or generate direct links with particular data sources in order to fully leverage integration with data sources.

8. BRINGING IT ALL TOGETHER: PROCTER & GAMBLE

There have been reports of the use of these technologies for supply network integration. Throughout, P&G has played an active role in extending supply network concepts. Anthes (2003) reported on how six different companies were using agent-based software for a wide range of tasks, including P&G. P&G was reported to save $300 million annually because of its ability to transform its supply chain. In addition, mirror world-type of capability was being reportedly studied by SAP, the large enterprise resource planning software firm, whose software was being used by P&G. Now a prototype, mirror world-like capabilities apparently are being built into SAPs supply chain software, to meet the needs of P&G and other advanced supply network firms.

Procter & Gamble's Modeling Supply Networks

Melymuka (2002) reported on P&Gs development of what appeared to be "mirror-world" capabilities: "companies are beginning to use complex adaptive systems to plot future business scenarios." P&G built an agent-based model to simulate their complex supply network. P&Gs model of their supply chain was designed to address questions, such as "What if supermarkets and other customers shared information about planned product promotions that might change their supply needs?" Using this model, and asking key questions supposedly has led to the finding of millions of dollars of potential savings.

According to Radjou, Orlov, and Nakashima (2002), the simulation allowed P&G to include in the model planning, sourcing, production, and delivery policies employed throughout the supply network. Agents were used to represent the many actors in the network. The modeling found counter intuitive results. Inventory could be decreased, stock-outs could be decreased and product could be sped through the network using so-called "less than truckload" (LTL) shipments and by combining multiple stock-keeping units (SKUs) in the same shipments. As reported in NuTech (2003) there was a 50% savings in cycle time and inventory, ultimately leading to a $300 million annual savings, on an investment of less than $3 million.

Procter & Gamble's Proactive Use of Intelligent Agents

Anthes (2003) provides a glimpse of P&Gs supply chain of the future, based on P&G and Forrester Research's example of the use of intelligent agents to proactively manage supply networks. In that vision, by the year 2008, P&G shortens the end-to-end cycle of replenishing a box of their detergent "Tide" from four months to one day. In the example, specialized production plants are replaced with flexi-plants and agents interact with each other and have a number of uses. First, intelligent agents monitor weather to determine when weather might impact the delivery of shipments, either by boat, train or truck. Second, intelligent agents are used to create alternative delivery schedules should they find problems with existing schedules due to issue such as the weather or problems with the delivery media, such as flat tires. Third, intelligent agents are used to bid for different production opportunities, based on availability particular production facilities. Fourth, intelligent agents gather real-time data from stores and warehouses, and use that data to estimate production requirements. Fifth, intelligent agents monitor shelves in stores to determine stocking needs, alerting stockers when to stock product on the shelves.

9. SUMMARY

This paper has examined and applied the use of mirror worlds in a supply network environment. It extended the concept of mirror worlds to be in the real world, suggesting that parts of the mirror world integrate with the "real world." If the mirror world capabilities are strong enough to really "mirror" the real world, then in many cases it is likely more effective to let the mirror world merge with the real world. The paper also illustrated the discussion with the analysis of a real world design by P&G, SAP, and BiosGroup.

Selected Web Addresses

BiosGroup	http://www.biosgroup.com/
SAP Supply Chain Management	http://www.sap.com/solutions/business-suite/scm/index.epx
Sante Fe Institute	http://www.santafe.edu

REFERENCES

Anthes, G. (2003). Agents of change. *Computerworld*, *27*(January), 26–27.

Cisco (1998). *Networking products marketplace*. http://www.cisco.com/public/ipcguest.html.

Croft, D. (1997). Intelligent software agents, http://www.alumnus.caltech.edu/~croft/research/agent/definition/.

Dstar (2001). SAP/BiosGroup team to solve supply network complexity, http://www.hpc-wire.com/dsstar/01/0619/103169.html.

Gelernter, D. (1992). *Mirror worlds: Or the day software puts the universe in a shoebox .... How it will happen and what it will mean?* New York, Oxford University Press, NY (first published in 1991).

Maloney, D. (2005). They've got to learn, DC Velocity, May, http://dcvelocity.com/articles/20050501/technologyreview.cfm.

McAfee, A., & McFarland, W. (2004). *Enterprise IT at Cisco* (*2004*), Harvard Business School, 9-605-015, November.

Melymuka, K. (2002). What if ... ? *Computerworld*, *4*(February), 26–27.

NuTech Solutions. (2003). Case study: The Procter & Gamble company.

Radjou, N., Orlov, L., & Nakashima, T. (2002). Adaptive agents boost supply network flexibility. *Forrester Research Report Brief*, 11 March.

SAP (2001). SAP evolves supply chains into adaptive supply chain networks, http://www11.sap.com/company/press/press.epx?pressID=179.

Thomas, S. (2002). *SAP Survey Project, SAP Corporate Research*, unpublished presentation, http://www.agentlink.org/agents-london/presentations/Bettina.pdf, Agents for Commercial Applications, January.

PART IV: FINANCIAL AND OTHER APPLICATIONS

This page intentionally left blank

AN ANALYTIC NETWORK PROCESS MODEL FOR INTERNAL AUDITOR SELECTION

Joseph Sarkis and Inshik Seol

ABSTRACT

There has been an increasing amount of research on personnel selection in many business disciplines (Hough & Oswald, 2000; Breaugh & Starke, 2000). Research on internal auditor selection, however, has had limited exposure in the auditing literature (Bailey, Gramling, & Ramamoorti, 2003). Recently, Seol and Sarkis (2005) introduced an analytic hierarchy process (AHP) model that used a decision hierarchy based on the CFIA (competency framework for internal auditing) framework. A limitation of AHP, however, is the assumption of strict hierarchical relationship that needs to exist among factors.

The purpose of this paper is an introduction of a more robust model, the analytical network process (ANP), which relaxes the strict hierarchical and decomposition levels of the hierarchy and incorporates possible interrelationships and interdependencies of various personnel selection criteria, factors, and alternatives. In illustrating the application, we return to the CFIA model framework, describe how and where interdependencies exist amongst the CFIA factors/attributes, and how ANP is used in the internal auditor selection process. The illustration will also describe some sensitivity analysis for the ANP approach. The tool is not without its

Applications of Management Science: In Productivity, Finance, and Operations
Applications of Management Science, Volume 12, 215–234

ISSN: 0276-8976/doi:10.1016/S0276-8976(06)12015-5

limitations that include the potential for geometrically more questions and information elicitation from the decision makers. Finally managerial and research implications associated with the technique and results are described.

INTRODUCTION

The selection of competent quality internal auditors is important in any organization since internal auditing plays a critical role in an organization to "evaluate and improve the effectiveness of risk management, control, and governance processes" (GTF Report, 1999). Even though there has been an increasing amount of research regarding personnel selection for organizations in other disciplines (Hough & Oswald, 2000; Breaugh & Starke, 2000), research in internal auditor selection has had limited exposure in the auditing research (Bailey et al., 2003). Recently, Seol and Sarkis (2005) introduced a multi-attribute technique called the analytic hierarchy process (AHP) to provide insight into how the technique can be applied to internal auditor selection. Their proposed model evaluated a hierarchy of skills and attributes internal auditors must possess to perform their tasks competently. These skills and attributes are based on the competency framework for internal auditing (CFIA) published by the Institute of Internal Auditors (IIA).

One of the limitations of the AHP model proposed by Seol and Sarkis, however, includes the strict hierarchical relationships among factors that need to be considered. According to CFIA, skills required to be competent internal auditors can be categorized as cognitive and behavioral skills that jointly affect internal auditors' performance. Cognitive skills are sub-categorized as technical skills, analytic/design skills, and appreciative skills; behavioral skills as personal skills, interpersonal skills, and organizational skills. These sub-category skills are also inter-related among each other. For example, personal skills are necessary for interpersonal performance, and together they are likely to be necessary for organizational performance. Cognitive skills have similar relationships (Birkett, Barbera, Leithhead, Lower, & Roebuck, 1999).

Due to its nature of interdependencies among skills recognized by CFIA and the psychological literature, an extension to Seol and Sarkis' model is to relax the strict hierarchical and decomposition levels of the hierarchy to more realistically incorporate possible interrelationships and interdependencies of the factors. This paper applies the 'non-linear' multi-attribute decision-making approach defined as the analytic network process (ANP).

The remainder of the paper is organized as follows: The next section discusses ANP and its appropriateness in internal auditor selection. A practical illustrative example will follow. Finally, a summary of the implications and limitations will be discussed along with future research streams.

THE ANALYTIC NETWORK PROCESS

Saaty (1980) first introduced AHP for decision structuring and decision analysis. One of its major advantages is allowing users to break down and categorize factors that can be ranked within groups (locally ranked) and amongst factors of all groups (globally ranked). It completes this process through a series of pair-wise comparisons. This two-at-a-time comparison is advantageous when simultaneously and explicitly considering numerous factors since a typical individual can evaluate 7 plus or minus 2 factors simultaneously (Miller, 1956).

AHP, however, assumes a unidirectional hierarchical relationship among decision levels. This simplistic assumption does not consider the many possible relationships among the groups of factors or those relationships within groups. For example, in selection of a project, a decision maker may categorize factors into cost, quality, and flexibility. A project may be rated on each of these factors separately and aggregated to arrive at an overall score, which is essentially what AHP does. Yet, AHP does not explicitly consider the interactions among the various factors (e.g. cost and quality may impact flexibility).

ANP is a more general and less frequently applied form of AHP, which is capable of incorporating interrelationships of factors into the decision model (Saaty, 1996). Saaty (1980) has also called this technique the ‘systems-with-feedback’ approach and argues that it is non-linear in how factors may be related to each other (interdependencies exist amongst and between levels).

Another problem with AHP is “rank reversal” that is defined as “the process of a change in rank ordering among already ranked alternatives on a set of criteria, when a new alternative is added to the group or an old one is deleted, and no additional criteria are introduced or deleted” (Saaty, 1996, p. 38). This has been a point of much debate as a theoretical and practical limitation of AHP (Dyer, 1990). This problem is mitigated with the ANP method making it a potentially more accurate and useful decision support tool for auditor selection (Schenkerman, 1994).

The network and non-linear structure of ANP is defined graphically with two-way arrows (or arcs), which represent interdependencies among clusters or groupings, or if within the same level of factors, a looped arc. The directions of the arcs signify dependence. Arcs emanate from a "controlling" attribute to other attributes that may influence it. The relative importance or strength of the impacts on a given element is measured on a ratio scale. The ANP approach is capable of handling interdependence among elements by obtaining the composite weights through the development of a "super-matrix". Overall, there are five major steps in the ANP process: (1) Develop a decision network hierarchy showing the relationships among decision factors. (2) Elicit pair-wise comparisons among the factors influencing the decision. (3) Calculate relative-importance-weight vectors of the factors from pair-wise comparisons. (4) Determine final relative importance weights of interdependent (network) factors using super-matrix. (5) Aggregate values of all candidates as desirability indices.

The Decision Network Hierarchy

The decision network hierarchy model of factors and relationships used to make the auditor selection decision is shown in Fig. 1.

We begin with the objective of selecting internal auditors. The two 'clusters' of factors are the general categories of cognitive and behavioral skills (Cluster Level). These skills are clustered into three sub-factor groupings each (Factor Level). The "Cognitive Skills" cluster incorporates technical, analytic/design, and appreciate skills factors. The "Behavioral Skills" cluster contains personal, interpersonal, and organizational skills factors. Table 1 defines these skills.

The "Specific Skills Level" can be quite numerous as shown in Table 2. They are representative of more specific internal auditor skills that an organization seeks and may be selected based on the level of auditor sought (i.e. entering level vs. more experienced internal auditors). The large number of specific skills factors may require some a priori screening of factors to be used in the evaluation. Empirical studies and evaluations of these skills, and factors using some form of clustering or factor analysis, may help to further define groupings and reduce skills dimensionality.

The "Candidate Level" of consideration includes the alternative candidates who are sought for the internal auditor position. It is at this stage that the performance of each of the candidates will be evaluated on the various specific skills factors. To help maintain a parsimonious model, we only consider interdependencies at the "Factors Level".

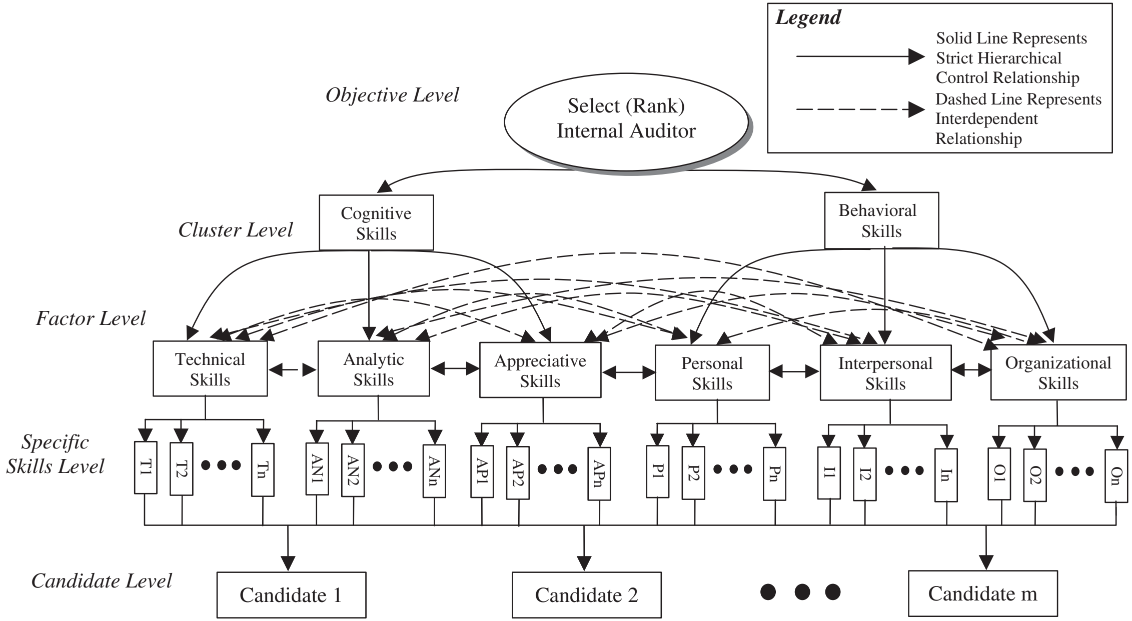


Fig. 1. General Network Decision Hierarchy for ANP Analysis with Objective of Selecting Internal Auditor.

Table 1. Skills Required of Internal Auditors.

Skills Required of Internal Auditors	
	Cognitive skills
Technical skills	Following defined routines with some mastery
Analytic/design skills	Problem identification or task definition and the structuring of prototype solutions or performances
Appreciative skills	Making complex and creative judgments, often in situations of ambiguity
	Behavioral skills
Personal skills	Handling oneself well in situations of challenge, stress, conflict, time pressure, and change
Interpersonal skills	Securing outcomes through interpersonal interactions
Organizational skills	Securing outcomes through the use of organizational networks

Source: Reproduced from Research Opportunities in Internal Auditing.

The interdependencies of behavioral and cognitive skills occur both within and between these clusters. Such interdependencies are recognized in the original study by the Institute of Internal Auditors (Birkett et al., 1999). As discussed before, individual attributes by internal auditors (i.e. cognitive and behavioral skills and elements within their cluster) jointly influence each other. For example, there has been psychological literature that shows personality (behavioral) and cognitive characteristics that relate to each other and as well as organizational and personal skills factors such as creativity and teamwork (James & Asmus, 2001; McClough & Rogelberg, 2003; Meneely & Portillo, 2005).

In many of these studies the strengths of these cross influences has been established, but is still heavily dependent on some of the task and social environments faced by individuals. This characteristic still leads to a dynamic relationship among the variables. Depending on the employee's responsibilities, organizational structure, task environment, etc., employee and selection characteristics will have varying importance levels. This issue of dynamic personnel selection criteria, and their relationships, has been well documented in the assessment and selection literature (Steele-Johnson, Osburn, & Pieper, 2000).

The dynamism in the selection criteria and their importance has significant implications for long-term suitability of selected employees, and may require consideration of temporal and inter-criteria relationships (e.g., the influence of learning and cognitive criteria over time may influence organizational skills learning). Thus, a model that is malleable in the relative

Table 2. Individual Attributes Required of Internal Auditors.

Cognitive Skills			Behavioral Skills		
Technical Skills	Analytic/Design Problem Structuring and Solving Skills	Appreciative Skills Judgment/Synthesis	Personal Skills	Interpersonal Skills	Organizational Skills
•Using information technology database systems spreadsheets •Communication literacy/writing structuring reports •Using relevant statistical methods •Understanding of organizational dynamics •Understanding of theories of risk •Understanding of theories of organizational control •Using information technology audit software •Apply control system designs and procedures •Documentation of internal audit work •Using/reviewing accounting procedures/principles •Apply laws and regulations •Apply internal auditing technologies and procedures •Using/reviewing accounting principles/procedures	•Logical reasoning •Ability to conceptualize •Problem analysis/structuring •Research skills (finding, accessing, and assessing data) •Using data in problem solving •Linking evidence to arguments and conclusions •Analyzing commercial and financial data •Basic analysis of accounts and accounting reports •Systems analysis and review •Internal audit requirements analysis/definition •Using sophisticated analytic models in support of internal audit judgment •Using industry-specific databases in the internal audit process •Using comprehensive internal auditing approaches	•Recognize importance of/in data •Sorting out the relevant (e.g., in data, evidence) •Judging whether information is sufficient, supportive of opinions •Observant/aware •Critical thinking •Able to be concise/succinct •Accepting of new/other's ideas •Strives for continuous improvement in self •Finding all that is relevant •Sorting out productive, central lines of inquiry •Seeing anomalies and recognizing their implications •Sensing the significance of issues •Seeing the internal audit process as a whole •Locating particular problems or situations in terms of more global contexts/responsibilities •Making associations, thinking outside the square	Honesty •Integrity •A-political •Inquisitive •Questioning •Balanced (not blinkered) in line of inquiry •Not dogmatic •Has initiative •A self-starter •Intelligence •Open minded •Flexible •Adapting to circumstances •Effecting change in self •Creativity •Objective •Positive attitude to technology •Sociable •Confident •Enthusiasm •Accepting of responsibility •Making it happen •Ability to handle pressure •Time management •Decisive •Able to stand ground •Stress management •Patience	•Communication awareness of audience listening oral interpersonal •Presentation skills •A team player •Learning from others •Handling frustration for self •Discretion/tact •Empathy •Culture sensitivity •Communication persuasiveness •Influencing, persuading, motivating, changing others •Leaderships of teams, groups •Handling an adversarial role •Handling multi-tasking •Able to diffuse conflict, conflict resolution •Ability to calm a situation •Handling frustration for others •Manage intra-group dynamics	•Finding way around organizations •Attaining a knowledge of the business (products, strategies, processes, markets, risks) •Adapting internal audit work to a wide range of organizational systems, methods, and standards •Negotiating the application of professional standards •Adding commercial value •Making productive gains •Marketing internal auditing services •Using sophisticated technologies/approaches in managing internal audit work TQM project management time management using performance criteria benchmarking planning •Scheduling

Table 2. (*Continued*)

Cognitive Skills			Behavioral Skills		
Technical Skills	Analytic/Design Problem Structuring and Solving Skills	Appreciative Skills Judgment/Synthesis	Personal Skills	Interpersonal Skills	Organizational Skills
•Master of new information technologies •Understanding key principles of specialty fields, including Environmental management systems quality management systems Information Technology controls	•Using extra-organizational information in the internal audit process •Using non-financial evaluation methods in internal audit work •Designing control systems •Organizational analysis Strategies Functions (financing, marketing, production) Structures Processes Risks Controls •Using models in analysis •Adapting internal audit methodologies for evaluating controls in computer systems •Validating assumptions/projections underpinning plans and decisions •Developing prototype solutions to problems •Development of technologies for reducing audit risk	•Comprehending internal audit in the context of business •Discriminating between substance and form •Disciplining imagination •Sensing/serving client needs and expectations •Having a sense of practicability, materiality •Assessing the risk associated with internal audit assignments •Coping with increasingly complex transactions, regulations and organizations •Adapting to revised expectations about internal audit processes and outcomes •Risk awareness •Interpreting relevant laws and standards •Applying disciplinary understandings and research findings to internal audit work •Extending judgment over time (projection)	•Persistence •Dedication •Intuitive/gut-feel •Tenacity •Determination •Handle/welcome change •Proactive •Assertive •Professional demeanor •Pushing the limits •Incisive •Anticipation	•Defines requirements (for team, others) •Securing control •Coaching/mentoring •Develop others •Manage inter-group dynamics •Delegation within teams •Conducts meetings •Facilitation •Liaison/negotiation within team for team	•Coping with international transactions, structures and legal arrangements •Building/using relationships and networks •Adding client value •Building trust •Managing internal audit work •Using organizational power sources, and structures •Delegations within function •Reading the culture and politics of an organization •Liaison and negotiation for function •Leadership of the function •Using consulting entrepreneurial approaches in selecting areas of internal audit work •Expanding internal audit work into new areas (requiring new skills)

	• Designing new internal audit technologies for systems analysis and evaluation • Developing internal audit technologies for assessing business risk • Developing methodologies and databases for establishing performance criteria and measuring performance • Design of risk management systems	• Knowing what should be there • Sensing what is not there • Making syntheses from isolated evidence • Employing a sense of perspective • Taking a strategic view, seeing the macro as well as micro • Being street wise (applying a sense of commercial reality) • Business acumen • Cope with information overload • Able to see big picture • Expect/cope with the bizarre • Desire for win-win • Making sense of complex situations • Managing complexity • Seeks to add value • Seeks to instill quality • Juggling inconsistent priorities • Strives for continuous improvement in others • Developing and using criteria to promote consistency in judgment • Making complex, multi-valued judgments in the absence of data and with probabilistic inferences only			• Human resource management • Strategy formation (for function) • Championing empowerment

Note: Entering Internal Auditor – Italics. Competent Internal Auditor – Normal Type. Internal Auditing Management – Bold Type.
Source: Reproduced from Research Opportunities in Internal Auditing.

importance of criteria for a given organization is certainly a characteristic that is required for effective evaluation. But, it is still left up to the decision maker to provide input into the relative importance of the different factors and the level of interdependencies, unless significant research exists to show interrelationships (e.g., through regression analysis of previous results).

The next step in ANP is to elicit preferences through pair-wise comparisons of the various factors and alternatives. This step is completed by asking a series of questions that compare the relative importance or influence of one factor (e.g. *Cognitive Skills*) when compared to another factor at the same level (i.e. *Behavioral Skills*) on a "controlling" factor (e.g. the selection of an internal auditor.).

These pair-wise comparisons are summarized in a pair-wise comparison matrix. One pair-wise comparison matrix will be formed for each of the levels and interdependencies and the respective influences of the factors. The illustrative example provides additional detail on this value elicitation step in the next section.

The third step in the ANP process is to complete the evaluations of the factors' relative importance weights by determining a local priority vector that is computed as the unique solution to:

$$Aw = \lambda_{\max} w \tag{1}$$

where $\lambda_{\max}$ is the largest eigenvalue of pair-wise comparison matrix A. Saaty (1980) provides several algorithms for approximating w, the final relative importance weights of the factors. The solution for the eigenvalues and relative importance weights in this paper is calculated using Web HIPRE3+ (Mustajoki & Hämäläinen, 2000), an Internet, interactive software decision support tool available for decision analysis (http://www.hipre.hut.fi/).

The fourth step in the process is the formation of a super-matrix. The super-matrix is composed of the relative importance weights from the network portion of the network decision hierarchy. This includes the relationships within and between "Factors Levels" skills including *technical*, *analytic/design*, *appreciative*, *personal*, *interpersonal* and *organizational* skills. To arrive at relative importance weights for evaluation purposes, we complete a Markovian-based analysis of the pair-wise comparison of relative importance scores in the matrix to achieve a stable or convergent set of weights. To do that, we first normalize the weights in each column to sum up to one (i.e. make the matrix "column stochastic"). Then we complete this task by dividing every element in a column by the sum of that column. We have designed the network to converge since the convergence to a stable set of weights is not always guaranteed (Saaty, 1996). To obtain the converged set of weights, we raise the super-matrix to a large power, that is

$$W_{\mathrm{F}} = \underset{p \to \infty}{\mathrm{Limit}} (W_{\mathrm{I}})^{p} \tag{2}$$

where $\boldsymbol{W}_{\mathrm{F}}$ and $\boldsymbol{W}_{\mathrm{I}}$ represent the final and initial super-matrices, respectively.

The final step of the process is to aggregate the values to arrive at the final scores for each candidate. This aggregation is a weighted average sum calculation defined by expression (3)

$$Auditor_i = \sum_{l=1}^{2} \sum_{k=1}^{K_t} \sum_{j=1}^{J_l} A_{ijkl} SS_{jkl} F_{kl}^{I} F_{kl}^{D} C_l \tag{3}$$

where

Auditor$_i$ is the overall desirability index (composite) score for *Auditor i.*

J_k is the index for the number of selected Specific Skills for a Factor k. This value may vary depending on the level of internal auditor that is being sought (e.g. there may be fewer skills required for an entry level internal auditor).

K_l the index for the number of selected Factors for a Cluster l. This value may also vary depending on the level of internal auditor that is being recruited (e.g. the *Organizational Skills* factor may not be required for an entry level internal auditor).

C_l the relative importance score of a Cluster l at the top level (e.g. a score for the *Cognitive Skills* cluster).

F_{kl}^{D} the direct (dependent) relative importance score of Factor skill k within a Cluster l (e.g. a score for *Technical Skills* factor which appears within the *Cognitive Skills* cluster).

F_{kl}^{I} the interdependent relative importance score of a Factor k within the Cluster l as determined by the super-matrix results.

SS_{jkl} the relative importance score for a Specific Skill j controlled by Factor k within Cluster l (e.g. a score or the *Communication* specific skill under the *Technical Skills* factor within the *Cognitive Skills* cluster).

A_{ijkl} the relative importance score of an Auditor candidate i for Specific Skill j under Factor k within the Cluster l.

Once this aggregation is complete a set of scores representing each of the candidates is then provided. A sensitivity analysis can also be completed to determine how value ranges for various factors will influence the rankings of the candidates. Our process of sensitivity analysis will vary scores of the relative importance of the *Cognitive* versus *Behavioral Skills* clusters. Sensitivity of one group of factors at a time can be completed for any level or set of factors by varying one factor's score, and making sure that the relative importance scores of the other factors remains at the same ratio among them.

ILLUSTRATIVE EXAMPLE

To further illustrate the ANP methodology and provide some insights into the process, we introduce an example with an assumption that we are seeking to select from among three candidates for a *competent level* internal auditor position (this is a level defined by CFIA to be between entry level and management level auditor positions). The first step requires us to be certain that the hierarchy is developed in such a way to represent this situation. Table 3 describes the specific skill elements within a hierarchy for competent level internal auditors.

When some of the factors overlap (e.g. honesty and integrity) we aggregated them into one factor (morality) by following the same groupings published by IIA (Birkett et al., 1999). Yet, it is up to the decision maker and organization to determine if they wish to use all or some of the factors, and whether aggregation of factors to simplify the model is an effective goal for them.

In this example there is a range of 3–8 factors for the specific skills groupings. Overall, for the whole decision network hierarchy, we will have 51 pair-wise comparison matrices ranging in size from 2×2 to 8×8. There will be a total of 275 pair-wise comparison questions required for this illustrative example.

To elicit the relative importance values, we need to develop pair-wise comparison questions. An example of pair-wise comparison question for the relative importance of *technical* versus *analytic* skills for an organization may be structured as thus: "How much more important are *technical skills* than *analytic skills* in contributing to the overall *cognitive skills* of a competent level internal auditor?" Similar questions will be asked for each pair-wise comparison relationship. For the lowest level of comparisons for the alternate candidates among one another, an example of pair-wise comparison question would be: "How much better is *Candidate 1* than *Candidate 2* at *Communication* (*T1*)?" The responses to these questions would be the 1/9 to 9 scale (from extremely less to extremely more influence or importance),

Table 3. Cognitive Skill Factors Used in Illustrative Example.

Technical Skills factors	Analytic/Design Skills Factors	Appreciative Skills Factors
Communication (T1)	Information literacy (AN1)	Discrimination (AP1)
Numeracy (T2)	Research (AN2)	Precision (AP2)
Computer Literacy (T3)	Problem structuring/ resolution (AN3)	Critique (AP3)
Organizational Understandings (T4)	Commercial and financial analyses (AN4)	Responsiveness (AP4)
Accounting/Financial Literacy (T5)	Organizational analysis (AN5)	Value Orientations (AP5)
Legal Literacy (T6)	Internal auditing applications (AN6)	Strategic Thinking (AP6)
Internal Auditing Application (T7)	Systems design (AN7)	Internal Auditing Approaches (AP7)
		Complexity Management (AP8)
Behavioral Skill Factors Used in Illustrative Example		
Personal Skills Factors	Interpersonal Skills Factors	Organizational Skills Factors
Morality (P1)	Communication (I1)	Organizational awareness (O1)
Inquisitiveness (P2)	People Skills (I2)	Value negotiation (O2)
Balance (P3)	Team Management (I3)	Task management (O3)
Flexibility (P4)		Function development (O4)
Directed (P5)		
Coping (P6)		
Intelligence (P7)		

recommended by Saaty (1980). The questions are also the same for interdependent relationships (same cluster/levels) within the decision network hierarchy.

A pair-wise comparison matrix example for the interdependent relationships among each of the "Skills Factor" level elements on the *Technical Skills Factor* is shown in Table 4. Across the first row of Table 4, one pair-wise comparison shows that the *Analytic Skills* factor has slightly more influence on and or relationship to the *Technical Skills* factor than the *Organizational Skills* factor (with a value of 6). Overall the relative importance weights column (w) shows the final relative scores (influences) for each of Skills factors on the *Technical Skills* factor. The results show that *Analytic Skills* have the most influence on *Technical Skills* overall (relative weight of 0.470), with *Personal*, *Appreciative*, *Interpersonal*, and *Organizational* Skills following with respective scores of 0.184, 0.171, 0.107, and 0.069.

Table 4. Pair-wise Comparison Matrix and Relative Importance Weight Results for Other Skills Factors and Impact on Technical Skills Factor.

Technical Skills	Analytic	Appreciative	Personal	Interpersonal	Organizational	*W*
Analytic	1	3	3	4	6	0.470
Appreciative	1/3	1	1	2	3	0.171
Personal	1/3	1	1	2	3	0.184
Interpersonal	1/4	1/2	1/2	1	2	0.107
Organizational	1/6	1/2	1/3	1/2	1	0.069

Table 5. Initial Super-matrix for Factors Interrelationships.

	Technical	Analytic	Appreciative	Personal	Interpersonal	Organizational
Technical	1.000	0.437	0.168	0.115	0.074	0.099
Analytic	*0.470*	1.000	0.145	0.166	0.112	0.092
Appreciative	*0.171*	0.110	1.000	0.299	0.246	0.198
Personal	*0.184*	0.246	0.384	1.000	0.317	0.216
Interpersonal	*0.107*	0.140	0.200	0.276	1.000	0.396
Organizational	*0.069*	0.066	0.103	0.144	0.251	1.000

Italic numbers are from Table 4.

After we have all the pair-wise comparisons completed, we go to our next step, to evaluate those factors with interdependencies using a super-matrix analysis. Since the results of Table 4's pair-wise comparisons are for an interdependent relationship, these relative importance scores will be included in the super-matrix. Table 5 shows this super-matrix. Notice that the italic numbers underneath the *Technical Skills* factor are from Table 4. Also note that the relative influence of *Technical Skills* on itself is 1.000, which we also assume for each of the other factors on themselves. The next step is to make the super-matrix shown in Table 5 column stochastic by dividing every element within each of the columns by the summation of that column (that is each column is divided by 2), before we raise the super-matrix to a large enough power to achieve convergence.

In our case, convergence (to the 4th decimal place) occurred when we raised the super-matrix to the 32nd power ($p = 32$). These converged values turn out to be $W_{\mathrm{F}} = 0.150, 0.163, 0.175, 0.217, 0.179, 0.115$ for *technical*, *analytic/design*, *appreciative*, *personal*, *interpersonal*, and *organizational skills*, respectively. The F_{kl}^{I} column of Table 6 shows these results.

Table 6. Desirability Index Calculations for Internal Auditor Selection/Ranking.

Cluster	Cl	Factor	F^D_{kl}	F^I_{kl}	Specific Skill	SS_{jkl}	A_{1jkl}	A_{2jkl}	A_{3jkl}	Auditor 1	Auditor 2	Auditor 3
Cognitive skills	0.667	Technical skills	0.387	0.150	T1	0.239	0.200	0.400	0.400	0.00185	0.00370	0.00370
	0.667		0.387	0.150	T2	0.073	0.455	0.312	0.233	0.00129	0.00088	0.00066
	0.667		0.387	0.150	T3	0.345	0.345	0.399	0.256	0.00461	0.00533	0.00342
	0.667		0.387	0.150	T4	0.114	0.234	0.415	0.351	0.00103	0.00183	0.00155
	0.667		0.387	0.150	T5	0.136	0.133	0.222	0.645	0.00070	0.00117	0.00340
	0.667		0.387	0.150	T6	0.051	0.650	0.088	0.262	0.00128	0.00017	0.00052
	0.667		0.387	0.150	T7	0.041	0.165	0.689	0.146	0.00026	0.00109	0.00023
	0.667	Analytic skills	0.443	0.163	AN1	0.128	0.075	0.333	0.592	0.00046	0.00205	0.00365
	0.667		0.443	0.163	AN2	0.299	0.103	0.216	0.682	0.00148	0.00311	0.00982
	0.667		0.443	0.163	AN3	0.043	0.637	0.105	0.258	0.00132	0.00022	0.00053
	0.667		0.443	0.163	AN4	0.094	0.200	0.600	0.200	0.00091	0.00272	0.00091
	0.667		0.443	0.163	AN5	0.293	0.117	0.268	0.614	0.00165	0.00378	0.00866
	0.667		0.443	0.163	AN6	0.049	0.200	0.600	0.200	0.00047	0.00142	0.00047
	0.667		0.443	0.163	AN7	0.094	0.158	0.082	0.761	0.00072	0.00037	0.00345
	0.667	Appreciative skills	0.169	0.175	AP1	0.109	0.095	0.25	0.655	0.00020	0.00054	0.00141
	0.667		0.169	0.175	AP2	0.281	0.333	0.334	0.333	0.00185	0.00185	0.00185
	0.667		0.169	0.175	AP3	0.037	0.345	0.399	0.256	0.00025	0.00029	0.00019
	0.667		0.169	0.175	AP4	0.078	0.500	0.250	0.250	0.00077	0.00038	0.00038
	0.667		0.169	0.175	AP5	0.263	0.400	0.400	0.200	0.00208	0.00208	0.00104
	0.667		0.169	0.175	AP6	0.042	0.133	0.222	0.645	0.00011	0.00018	0.00053
	0.667		0.169	0.175	AP7	0.078	0.650	0.088	0.262	0.00100	0.00014	0.00040
	0.667		0.169	0.175	AP8	0.112	0.165	0.689	0.146	0.00036	0.00152	0.00032

Table 6. *(Continued)*

Cluster	Cl	Factor	F_{kl}^{D}	F_{kl}^{I}	Specific Skill	SS_{jkl}	A_{1jkl}	A_{2jkl}	A_{3jkl}	Auditor 1	Auditor 2	Auditor 3
Behavioral skills	0.333	Personal skills	0.540	0.217	P1	0.065	0.075	0.333	0.592	0.00019	0.00084	0.00150
	0.333		0.540	0.217	P2	0.065	0.103	0.216	0.682	0.00026	0.00055	0.00173
	0.333		0.540	0.217	P3	0.065	0.333	0.334	0.333	0.00084	0.00085	0.00084
	0.333		0.540	0.217	P4	0.191	0.286	0.571	0.143	0.00213	0.00426	0.00107
	0.333		0.540	0.217	P5	0.113	0.500	0.250	0.250	0.00220	0.00110	0.00110
	0.333		0.540	0.217	P6	0.191	0.400	0.400	0.200	0.00298	0.00298	0.00149
	0.333		0.540	0.217	P7	0.311	0.540	0.163	0.279	0.00655	0.00198	0.00339
	0.333	Interpersonal skills	0.163	0.179	I1	0.558	0.387	0.443	0.169	0.00210	0.00240	0.00092
	0.333		0.163	0.179	I2	0.122	0.540	0.297	0.163	0.00064	0.00035	0.00019
	0.333		0.163	0.179	I3	0.320	0.333	0.334	0.333	0.00104	0.00104	0.00104
	0.333	Organizational skills	0.279	0.115	O1	0.109	0.075	0.333	0.592	0.00009	0.00039	0.00069
	0.333		0.279	0.115	O2	0.485	0.103	0.216	0.682	0.00053	0.00112	0.00353
	0.333		0.279	0.115	O3	0.297	0.637	0.105	0.258	0.00202	0.00033	0.00082
	0.333		0.279	0.115	O4	0.109	0.200	0.600	0.200	0.00023	0.00070	0.00023
Desirability indices ($Auditor_i$)										0.0465	0.0537	0.0656

Table 6 also represents the aggregation of all the results to determine the desirability indices of each auditor candidate. The final result for each of the three auditors appears in the final row, last three columns. In this illustrative example, we see that auditor candidate 3 is most preferable and highest ranked with a value of 0.0656, followed by auditor candidate 2 (0.0537), then auditor candidate 1 (0.0465).

SENSITIVITY ANALYSIS

An issue that faces managers who would wish to apply this technique is the relative significance of the differences in the desirability indices. If the results are very close, managers may wish to introduce factors that were not considered in this evaluation (e.g. other factors, salary requirements, experiences, reference quality, etc.). Another method would be to complete a sensitivity analysis to determine solution robustness.

For example, we can vary the relative importance of the Cluster level criterion of *Cognitive Skills*. Fig. 2 shows this sensitivity analysis over a range of 0–1 relative importance level for *Cognitive Skills*. The current situation has perceived relative importance of *Cognitive Skills* at 0.667. By increasing the relative importance of *Cognitive Skills*, we see that auditor candidate 3 increases in desirability index ranking, while that of auditor candidate 1 decreases. At a relative importance weight of approximately 0.224 for *Cognitive Skills* and 0.776 for *Behavioral Skills*, candidate 3 and 1 are essentially equal in desirability. It seems that candidate 2 will always be less preferable than either candidate 1 or 3 no matter what relative importance score values are used for the cluster level items. Managerial decision makers may wish to note these sensitivities, especially in circumstances where agreement on the relative importance may diverge greatly.

SUMMARY, CONCLUSIONS AND IMPLICATIONS

The paper introduces and applies a non-linear multi-attribute model (ANP) in the internal auditor selection process. The ANP relaxes the strict hierarchical and decomposition levels of the hierarchy which is the limitation of AHP model proposed by Seol and Sarkis (2005). ANP incorporates possible interrelationships and interdependencies of various personnel selection criteria, factors, and alternatives. In illustrating the application, we used the CFIA framework published by IIA.

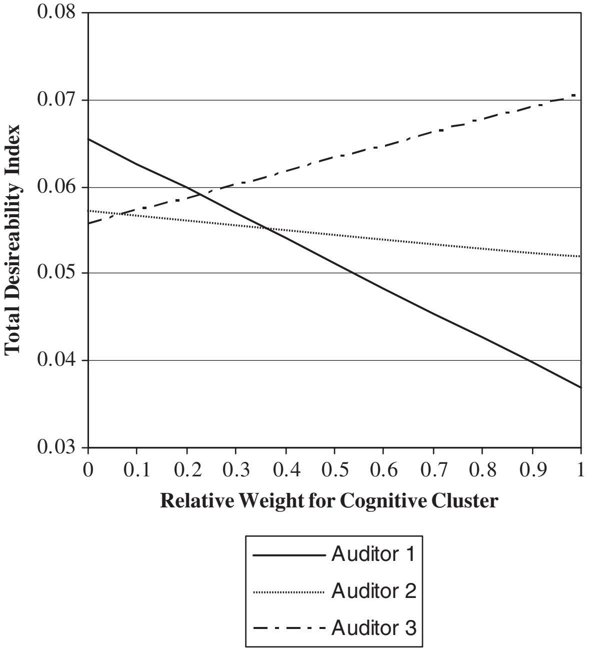


Fig. 2. Sensitivity Analysis for ANP Auditor Selection Problem for Cognitive Cluster Relative Importance Weight Ranges.

Even though ANP is a robust approach, it also has some limitations and concerns. First, the ANP approach may be disadvantageous in some situations because it can become quite complex as the number of factors and relationships increase, which requires more effort by analysts and decision makers. It is particularly troublesome if there happens to be great disagreement on the importance of factors.

Second, as identified by the Society for Industrial and Organizational Psychology (SIOP) (2003) care must be taken on the validation and use of appropriate weighting techniques for personnel selection. For example, the ANP process relies on decision makers' perceptual and managerial information, but such reliance does not exclude the use of criterion weights that could have been determined from previous studies using various statistical approaches such as regression analysis (Robertson & Smith, 2001). That is, previous performance results of workers using various predictors or criteria can be regressed to determine what weights are most appropriate for an ANP analysis.

Another issue in the ANP process arises when some factors with a larger set of sub-factors may have a diluted importance when compared to factors that have fewer sub-factors. Thus, a balance of factors or awareness of this characteristic needs to be explicitly stated and evaluated, but could be mitigated by inclusion of more interdependent relationships.

Some of the difficulties associated with the large amount of questions associated with the ANP approach may be overcome if the resultant weights for various elements are assumed to be relatively static (i.e. they do not change much over a lengthy period of time). In this situation, a one time initial effort of determining a baseline set of weights for the skills can be completed, and the only additional data requirements would be candidates' relative performance on these skills. That is, with repeated use of the model, the level of effort would tend to decrease because some weights would not change for a given organization.

The implications of this paper, however, are not restricted to selection of internal auditors. The same model can be used for internal auditor performance management and measurement purposes. This would be a post-selection audit of the performance of internal auditors on different factors. This post-selection information may also be used to update which factors should be used, the relative weightings, and the relationships amongst the selected factors.

Finally, we believe that actual application of the proposed model in a real setting will provide practical insight into the proposed ANP model. It is a slightly more complex application than the AHP model, but we believe managers grasp the process after minimal discussion and presentation.

REFERENCES

Bailey, A. D., Jr., Gramling, A. A., & Ramamoorti, S. (Eds) (2003). *Research opportunities in internal auditing*. Altamonte Springs, FL: Institute of Internal Auditors Research Foundation.

Birkett, W. P., Barbera, M. R., Leithhead, B. S., Lower, M., & Roebuck, P. J. (1999). *Competency: Best practices and competent practitioners*. Altamonte Springs, FL: Institute of Internal Auditors Research Foundation.

Breaugh, J. A., & Starke, M. (2000). Research on employee recruitment: So many studies, so many remaining questions. *Journal of Management*, *26*(3), 405–434.

Dyer, J. S. (1990). Remarks on the analytic hierarchy process. *Management Science*, *36*(3), 249–268.

GTF Report. (1999). *A vision for the future: Professional practices framework for internal auditing*. Report of the Guidance Task Force to The IIA's Board of Directors, The Institute of Internal Auditors, Altamonte Springs, FL.

Hough, L. M., & Oswald, F. L. (2000). Personnel selection: Looking toward the future – remembering the past. *Annual Review of Psychology*, *51*, 631–664.

James, K., & Asmus, C. (2001). Personality, cognitive skills, and creativity in different life domains. *Creativity Research Journal*, *13*(2), 149–159.

McClough, A. C., & Rogelberg, S. G. (2003). Selection in teams: An exploration of the teamwork knowledge, skills, and ability test. *International Journal of Selection and Assessment*, *11*(1), 56–66.

Meneely, J., & Portillo, M. (2005). The adaptable mind in design: Relating personality, cognitive style, and creative performance. *Creativity Research Journal*, *17*(2/3), 155–166.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *The Psychological Review*, *63*, 81–97.

Mustajoki, J., & Hämäläinen, R. P. (2000). Web-HIPRE – global decision support by value-free and AHP synthesis. *INFOR Journal*, *38*(3), 208–220.

Robertson, I. T., & Smith, M. (2001). Personnel selection. *Journal of Occupational and Organizational Psychology*, *74*, 441–472.

Saaty, T. L. (1980). *The analytic hierarchy process: Planning, priority setting, resource allocation*. New York: McGraw-Hill.

Saaty, T. L. (1996). *Decision making with dependence and feedback: The analytic network process*. Pittsburgh, PA: RWS Publications.

Schenkerman, S. (1994). Avoiding rank reversal in AHP decision-support models. *European Journal of Operational Research*, *74*, 407–419.

Seol, I., & Sarkis, J. (2005). A multi-attribute model for internal auditor selection. *Managerial Auditing Journal*, *20*(8), 876–892.

Society for Industrial and Organizational Psychology. (2003). *Principles for the validation and use of personnel selection procedures* (4th ed.). Ohio: Bowling Green.

Steele-Johnson, D., Osburn, H. G., & Pieper, K. F. (2000). A review and extension of current models of dynamic criteria. *International Journal of Selection and Assessment*, *8*(3), 110–136.

A TWO-DIMENSIONAL MINIMAL-REPAIR WARRANTY POLICY

Amitava Mitra and Jayprakash G. Patankar

ABSTRACT

A warranty policy involving two attributes, for example time and usage, is considered. Usage is assumed to be related to time through the usage rate, which is considered to be a random variable satisfying a specified probability distribution. The paper analyzes a policy where warranty is not renewed on product failure, within the specified time period and amount of usage, but is minimally repaired. Unit cost of minimal repair, conditional on the usage rate, is assumed to be a non-linear function of the two warranty parameters. Expressions for the expected warranty costs per unit sales are derived. Applications of the results are presented through sample computations. The results demonstrate the use of warranty cost information in selecting the parameters of the warranty policy.

INTRODUCTION

A majority of consumer products provide some sort of assurance to the consumer regarding the quality of the product sold. This assurance, in the form of a warranty, is offered at the time of sale. The Magnuson–Moss Warranty Act of 1975 (U.S. Federal Trade Commission Improvement Act,

Applications of Management Science: In Productivity, Finance, and Operations
Applications of Management Science, Volume 12, 235–251

ISSN: 0276-8976/doi:10.1016/S0276-8976(06)12016-7

1975) also mandates that manufacturers must offer a warranty for all consumer products sold for more than $15. The warranty statement assures consumers that the product will perform its function to their satisfaction up to a given amount of time (i.e., warranty period) from the date of purchase. Manufacturers offer many different types of warranties to promote their products. Thus, warranties have become a significant promotional tool for manufacturers. Warranties also limit the manufacturers' liability in the case of product failure beyond warranty period.

Manufacturers use warranties as a competitive strategy to boost their market share, profitability, and image. The cost of a warranty program must be estimated precisely and its effect on the firm's profitability must be studied. Manufacturers plan for warranty costs through the creation of a fund for warranty reserves. These funds are set aside at the beginning of the sales period to meet product replacement or repair obligations that occur while the product is under warranty. An estimate of the expected warranty costs is thus essential for management to plan for warranty reserves. For the warranty policy considered, we assume that the product will be minimally repaired if failure occurs within a specified time and the usage is less than a specified amount. Such a two-dimensional policy is found for products such as automobiles where the warranty coverage is provided for a time period, say 5 years, and a usage limit of, say, 50,000 miles.

In this paper we consider estimation of expected warranty costs for a two-dimensional warranty policy where the warranty parameters, for example, could be time and usage at the point of product failure. A warranty policy in this context, such as those offered for automobiles, could be stated as follows: Product will be replaced or repaired free of charge up to a time (W) or up to a usage (U), whichever occurs first from the time of the initial purchase. Warranty is not renewed on product failure. For example, automobile manufacturers may offer a 36 months or 36,000 miles warranty, whichever occurs first. For customers with high usage rates, the 36,000 miles may occur before 36 months. On the contrary, for those with limited usage, the warranty time period of 36 months may occur first.

We assume that usage is related to time as a linear function through the usage rate. To model a variety of consumers, usage rate is assumed to be a random variable with a specified probability distribution. The paper develops a model based on minimal-repair of failed items. This implies that the failure rate of the repaired items is the same as it was just prior to failure. While prior research has assumed the cost of repair to be independent of age and total usage at failure, in this paper the unit cost of minimal repair is assumed to be a nonlinear function of the age of the product and the usage at the time of failure.

We derive expressions for the expected warranty costs per unit sales. Some applications of the derived results are demonstrated in the managerial context. Procedures for the selection of warranty parameters are demonstrated based on financial restrictions imposed upon management.

Literature Review

Two-dimensional warranties are characterized by a region in a two-dimensional plane with one axis representing time (or age) and the other representing a second attribute, for example, usage. Item failures are events that occur randomly on this two-dimensional plane. Two approaches have been used to model such failures (Blischke & Murthy, 1994, 1996). In the first approach, usage is modeled as a function of time, which reduces the model to a one-dimensional point process formulation. In the second, failures are modeled by a two-dimensional point process formulation.

The majority of past research has dealt with a single-attribute warranty policy, where the warranty parameter is typically the time since purchase of the product. Singpurwalla (1987) developed an optimal warranty policy based on maximization of expected utilities involving both profit and costs. A bivariate probability model involving time and usage as warranty criteria was incorporated. One of the first studies among two-dimensional warranty policies using a one-dimensional approach is that by Moskowitz and Chun (1988). Product usage was assumed to be a linear function of the age of the product. Moskowitz and Chun (1994) used a Poisson regression model to determine warranty costs for two-dimensional warranty policies. They assumed that the total number of failures is Poisson distributed whose parameter can be expressed as a regression function of age and usage of a product. Murthy, Iskander, and Wilson (1995) used several types of bivariate probability distributions in modeling product failures as a random point process on the two-dimensional plane and considered free-replacement policies. Eliashberg, Singpurwalla, and Wilson (1997) considered the problem of assessing the size of a reserve needed by the manufacturer to meet future warranty claims in the context of a two-dimensional warranty. They developed a class of reliability models that index failure by two scales, such as time and usage. Usage is modeled as a covariate of time. Gertsbakh and Kordonsky (1998) reduced the two-dimensional problem to a single dimension by selecting a linear relationship between usage and time. Ahn, Chae, and Clark (1998) also present a unidimensional approach that combines the two indices into a single index using logarithmic transformation. Singpurwalla and Wilson (1998) develop probabilistic models indexed by two scales. The relationship between the scales of time and usage is described by an additive hazards model. Usage is described by stochastic processes such as the Poisson process.

Chun and Tang (1999) found warranty costs for a two-attribute warranty model by considering age and usage of the product as warranty parameters. They provided warranty cost estimation for four different warranty policies: rectangular, L-shaped, triangular, and iso-cost, and performed sensitivity analysis on discount rate, usage rate, and warranty terms to determine their effects on warranty costs. Kim and Rao (2000) considered a two-attribute warranty model for non-repairable products using a bivariate exponential distribution to explain term failures. Analytical expressions for warranty costs are derived using Downtone's bivariate distribution. They demonstrate the effect of correlation between usage and time on warranty costs. A two-dimensional renewal process is used to estimate warranty costs. Hsiung and Shen (2001) considered the effect of warranty costs on optimization of the Economic Manufacturing Quality (EMQ). As a process deteriorates over time, it produces defective items that incur reworking costs (before sale) or warranty repair costs (after sale). The objective of their paper is to determine the lot size that will minimize total cost per unit of time. The total cost per unit of time includes set up cost, holding cost, inspection cost, reworked cost, and warranty costs. Sensitivity analysis is performed on various costs to determine an optimum production lot size.

Yeh and Lo (2001) explored the effect of preventive maintenance actions on expected warranty costs. A model is developed to minimize such costs. Providing a regular preventive maintenance within the warranty period increases maintenance cost to the seller, but the expected warranty cost is significantly reduced. An algorithm is developed that determines an optimal maintenance policy. Lam and Lam (2001) developed a model to estimate expected warranty costs for a warranty that includes a free repair period and an extended warranty period. Consumers have an option to renew warranty after the free repair period ends. The choice of consumers has a significant effect on the expected warranty costs and determination of optimal warranty policy.

MODEL DEVELOPMENT

The following notation is used in the paper:

W	warranty period offered in warranty policy
U	usage limit offered in warranty policy
R	usage rate
t	instant of time
$Y(t)$	usage at time t
$\lambda(t\|r)$	failure intensity function at time t given $R = r$
$G(\mathrm{W},\mathrm{U}\|r)$	expected warranty cost for minimal repair per unit given $R = r$
p	Unit product sales price
$D(t\|r)$	cost of minimal repair per unit if product fails at age t given $R = r$
$C(\mathrm{W},\mathrm{U})$	expected warranty costs for minimal repair per unit

Relationships between Warranty Attributes

We assume that the two warranty attributes, say time and usage, are related linearly through the usage rate, which is a random variable. Denoting $Y(t)$ to be the usage at time t and $X(t)$ the corresponding age, we have:

$$Y(t) = RX(t) \tag{1}$$

where R is the usage rate. It is assumed that all items that fail within the prescribed warranty parameters are minimally repaired and that the repair time is negligible. In this context, $X(t) = t$.

Distribution Function of Usage Rate

In order to model a variety of customers, R is assumed to be a random variable with probability density function given by $g(r)$. The following distribution functions of R are considered in the paper.

(a) *R has a uniform distribution over (a, b)*. This models a situation where the usage rate is constant across all customers. The density function of R is given by:

$$\begin{aligned} g(r) &= \frac{1}{b-a}, \quad && a \le r \le b \\ &= 0, && \text{otherwise} \end{aligned} \tag{2}$$

(b) *R has a gamma distribution function*. This may be used for modeling a variety of usage rates among the population of buyers. The shape of the gamma distribution function is influenced by the selection of its parameters. When the parameter, p, is equal to 1, it reduces to the exponential distribution. The density function is given by

$$g(r) = \frac{e^{-r} r^{p-1}}{\Gamma(p)}, \qquad 0 \le r < \infty, \quad p > 0 \tag{3}$$

Failure Rate

Failures are assumed to occur according to a Poisson process where it is assumed that failed items are minimally repaired. If the repair time is small, it can be approximated as being zero. Since the failure rate is unaffected by minimal repair, failures over time occur according to a nonstationary Poisson process with intensity function $\lambda(t)$ equal to the failure rate.

Conditional on the usage rate $R = r$, let the failure intensity function at time t be given by

$$\lambda(t|r) = \theta_0 + \theta_1 r + (\theta_2 + \theta_3 r)t \tag{4}$$

(a) *Stationary Poisson process*. Under this situation, the intensity function $\lambda(t|r)$ is a deterministic quantity as a function of t when $\theta_2 = \theta_3 = 0$. This applies to many electronic components that do not deteriorate with age and failures are due to pure chance. The failure rate in this case is constant.

(b) *Nonstationary Poisson process*. This models the more general situation where the intensity function changes as a function of t. It is appropriate for products and components with moving parts where the failure rate may increase with time of usage. In this case θ_2 and θ_3 are not equal to zero.

Minimal-Repair Cost Function

The unit cost of minimal repair is assumed to be a function of the age, t, and usage, $Y(t)$, of the product at the time of failure, conditional on the usage rate. Such a cost function is justified for intricate products where wear and tear of mechanical parts take place as a function of usage of the product. Since usage has been assumed to be related to product age through the usage rate, the minimal-repair cost function can therefore be expressed as a function of the product age.

The selected form of the minimal-repair cost function is motivated by the concept of the Taguchi loss function (Taguchi, 1986; Taguchi, Elsayed, & Hsiang, 1989). Such a loss function increases quadratically with increasing deviation of the quality characteristic from a target value. Since wear and tear of mechanical parts increase with usage, the cost of minimal repair is assumed to be a quadratic function of usage, along these lines. While products deteriorate with age, too, just possessing the product may not necessarily indicate use of the product, which causes the wear of mechanical parts. Using these guidelines, the conditional unit minimal-repair cost function is given by

$$\begin{aligned} D(t|r) &= f + gt + h[Y(t)]^2 \\ &= f + gt + hr^2t^2 \end{aligned} \tag{5}$$

where f, g, and h are constants.

Selection of the minimal-repair cost function parameters f, g, and h could be based on historical cost data and product attributes at the time of failure. A least-squares multiple regression analysis may be used to obtain estimates of f, g, and h.

A deterministic approach may use repair costs that are known to be associated with certain predetermined values of age and usage. In this approach, let f denote the fixed cost of unit repair and let c denote the cost of a new item. The fixed cost is usually a certain fraction of the unit cost of the item, i.e.,

$$f = kc \tag{6}$$

where k is a constant. For example, k could be 0.4, if 40% of the unit cost is considered to be the fixed cost component. Suppose, the manufacture also estimates that the unit cost of repair is about the same as the cost of a new item for items that have age and usage combinations as $(W/2, 2U)$ and $(W, 5U/4)$. The implication of this is that for a product that fails at half the warranty period but twice the specified usage limit, the minimal cost of repair is c. Expressing the above statement in equation form, we have

$$c = f + gW/2 + h(2U)^2 \tag{7}$$

and

$$c = f + gW + h(5U/4)^2 \tag{8}$$

Simultaneous solutions of Eqs. (6)–(8) will lead to evaluation of the constants f, g, and h. This will lead to the determination of $D(t|r)$ in Eq. (5).

Note that the above approach is a suggested method of estimation of the parameters of the unit repair cost function. The procedure is also applicable if the manufacturer has an estimate of the unit minimal-repair cost at three different points in the two-dimensional space of time and usage. Thus, the minimal cost of repair need not be c, the cost of a new item. Consider, for example, the repair cost estimates to be c_1, c_2, and c_3 at the points (t_1, Y_1), (t_2, Y_2), and (t_3, Y_3), respectively. This information can then be used to obtain estimates of the cost function parameters of equation (5) through simultaneous solutions.

Expected Warranty Costs

The warranty region is the rectangle shown in Fig. 1, where W is the warranty period and U is the usage limit. Let $\gamma_1 = U/W$. Conditional on the usage rate $R = r$, if the usage rate $r \geqslant \gamma_1$, warranty ceases at time X_r, given by

$$X_r = U/r \tag{9}$$

Alternatively, if $r < \gamma_1$, warranty ceases at time W. Let $G(W, U|r)$ denote the expected warranty costs for minimal repair per unit conditional on $R = r$.

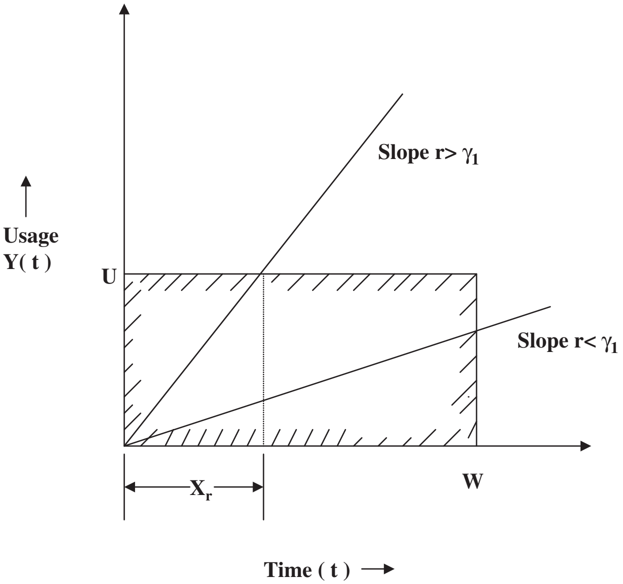


Fig. 1. Two-Dimensional Warranty Region.

We have

$$G(W, U|r) = \begin{cases} \int_{t=0}^{W} \lambda(t|r)D(t|r)\,\mathrm{d}t, & \text{if } r < \gamma_1 \\ \int_{t=0}^{X} \lambda(t|r)D(t|r)\,\mathrm{d}t, & \text{if } r \geq \gamma_1 \end{cases} \tag{10}$$

The expected warranty costs per unit for minimal repair is thus obtained from:

$$\begin{aligned} C(W, U) = & \int_{r=0}^{\gamma_1} \left[\int_{t=0}^{w} \lambda(t|r)D(t|r)dt \right] g(r)\,\mathrm{d}r \\ & + \int_{r=\gamma_1}^{\infty} \left[\int_{t=0}^{X_r} \lambda(t|r)D(t|r)dt \right] g(r)\,\mathrm{d}r \end{aligned} \tag{11}$$

The unit sales price, p, is usually based on the unit cost of the item and is given by:

$$p = mc \tag{12}$$

where m is a constant. Possible values of m could be, for example, 1.1 or 1.20.

Of interest to the manufacturer is the expected warranty costs per unit sales given by:

$$CS(W, U) = C(W, U)/p \tag{13}$$

Values of $CS(W,U)$ assist management in budgeting for warranty costs in the future.

RESULTS

To demonstrate application of the formulated model, sample results are obtained using selected values of the model parameters. The distribution of usage rate, R, is chosen to be uniform in the interval (0,6) and gamma with parameter $p = 4$. The failure rate intensity function, conditional on R, is selected to be for a stationary Poisson process with parameters $\theta_2 = 0$, $\theta_3 = 0$, $\theta_0 = 0.005$, and $\theta_1 = 0.01$. For the minimal-repair cost function, it is assumed that the unit repair cost is c, the cost of a new item, for the following age and usage combinations, respectively: $(W/2, 2U)$ and $(W, 5U/4)$. The fixed cost component parameter, k, is chosen as 0.4. Values of the sales price parameter, m, are selected as 1.1 and 1.2. Warranty parameter U denoting the usage limit is selected to be $U = 2.4$, 4.0, and 6.0 (for example 24,000, 40,000, and 60,000 miles), while the warranty time W is varied within the interval $W = (0.4, 5.0)$, i.e., 0.4–5 years, to study the impact on warranty costs.

Since closed-form expressions for expected warranty costs per unit are not easily obtainable for the different distributions of the usage rate, numerical integration techniques are used in the computations. In particular, subroutine QDAG of the International Mathematical and Statistical Library (IMSL) (1987) is used, which integrates a function using a globally adaptive scheme based on Gauss–Konrad rules.

Table 1 shows the expected warranty costs per unit sales for uniform distribution of the usage rate. The values can be used by management for budgeting and planning purposes as well as for selecting warranty parameter values. Observe that if the warranty period is set at 1.2 years, for $m = 1.2$, the expected warranty costs to the manufacturer is 15.02 cents per sales dollar, if the usage limit is set at 2.4 (or equivalently 24,000 miles).

Note that if management wishes to double the warranty period (for 1.2–2.4), all other conditions being the same, expected warranty costs would be 15.24 cents per sales dollar, an increase of about 1.5%. While this provides an estimate on the increase in expenditures in deciding on a value of the warranty policy parameter, another factor that decision makers must incorporate is the impact on sales as the warranty period is increased. It is

Table 1. Expected Warranty Costs Per Unit Sales for Uniform Distribution of the Usage Rate.

W	$U = 2.4$		$U = 4.0$		$U = 6.0$	
	$m = 1.1$	$m = 1.2$	$m = 1.1$	$m = 1.2$	$m = 1.1$	$m = 1.2$
0.4	0.1583	0.1451	0.2445	0.2242	0.3823	0.3504
0.8	0.1621	0.1486	0.2606	0.2389	0.4035	0.3699
1.2	0.1638	0.1502	0.2667	0.2445	0.4123	0.3779
1.6	0.1649	0.1512	0.2703	0.2477	0.4176	0.3828
2.0	0.1656	0.1518	0.2725	0.2498	0.4214	0.3862
2.4	0.1663	0.1524	0.2744	0.2515	0.4243	0.3389
2.8	0.1669	0.1530	0.2757	0.2528	0.4267	0.3912
3.2	0.1674	0.1534	0.2769	0.2538	0.4287	0.3930
3.6	0.1678	0.1538	0.2778	0.2547	0.4305	0.3947
4.0	0.1683	0.1543	0.2787	0.2555	0.4321	0.3961
4.4	0.1686	0.1546	0.2794	0.2562	0.4336	0.3975
4.8	0.1690	0.1549	0.2802	0.2568	0.4350	0.3988
5.0	0.1692	0.1551	0.2805	0.2571	0.4356	0.3993

Note: $a = 0$, $b = 6$, $\theta_0 = 0.005$, $\theta_1 = 0.01$, $k = 0.4$.

possible that the increase in sales in doubling the warranty period from 1.2 to 2.4 may more than offset the increase in warranty expenditures.

One can also observe the trade-offs in selecting the two warranty parameters, W and U, from Table 1. Observe that if the warranty period is selected as 2.0 years, for $m = 1.2$, as the usage limit increases from 2.4 to 6.0, the warranty costs increase from 15.18 to 38.62 cents per sales dollar. This is an increase of about 154%. Management has to identify sources of funds to budget such expenditures. Given resource constraints, this may involve reallocation of funds from other areas.

Fig. 2 shows graphs of expected warranty costs per unit sales, when the usage rate distribution is uniform, as a function of the warranty period W for values of the usage limit U being 2.4, 4.0, and 6.0, respectively, and $m = 1.2$. While the expected warranty costs per unit sales increase with W, for a given value of U, the rate of increase decreases with an increase in the warranty period. The graphs show a similar behavior for the selected values of U. As expected, the graphs for large values of U dominate those for smaller values of U, for a given W. Unit warranty costs become fairly constant for values of the warranty period of 3 or more years. Therefore, for a given usage limit, management may explore the added benefits from increased sales if the warranty period is extended say from 3 to 4 years.

Table 2 shows expected warranty costs per unit sales when the usage rate has a gamma distribution with parameter $p = 4$. Unit warranty costs are

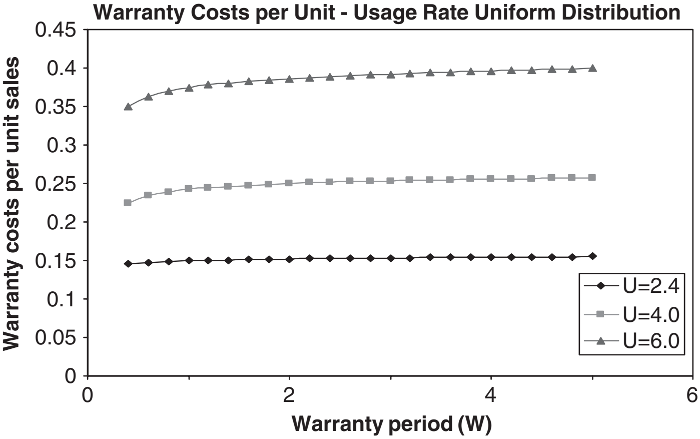


Fig. 2. Warranty Costs Per Unit Sales for Usage Rate having Uniform Distribution.

Table 2. Expected Warranty Costs Per Unit Sales for Gamma Distribution of the Usage Rate.

W	$U = 2.4$		$U = 4.0$		$U = 6.0$	
	$m = 1.1$	$m = 1.2$	$m = 1.1$	$m = 1.2$	$m = 1.1$	$m = 1.2$
0.4	0.0024	0.0022	0.0095	0.0008	0.0007	0.0007
0.8	0.0092	0.0084	0.0066	0.0061	0.0032	0.0029
1.2	0.0126	0.0116	0.0143	0.0131	0.0106	0.0097
1.6	0.0144	0.0132	0.0195	0.0179	0.0192	0.0176
2.0	0.0155	0.0142	0.0228	0.0209	0.0263	0.0241
2.4	0.0162	0.0148	0.0250	0.0229	0.0317	0.0291
2.8	0.0167	0.0153	0.0265	0.0243	0.0358	0.0328
3.2	0.0173	0.0158	0.0277	0.0254	0.0388	0.0356
3.6	0.0176	0.0162	0.0286	0.0262	0.0412	0.0377
4.0	0.0180	0.0165	0.0293	0.0268	0.0431	0.0395
4.4	0.0184	0.0168	0.0299	0.0274	0.0445	0.0407
4.8	0.0188	0.0172	0.0304	0.0278	0.0457	0.0419
5.0	0.0189	0.0173	0.0306	0.0281	0.0463	0.0424

Note: $p = 4$, $\theta_0 = 0.005$, $\theta_1 = 0.01$, $k = 0.4$.

much lower than the case when usage rate is uniformly distributed for the same values of the warranty parameters of time and usage. For instance, for $m = 1.2$, and $U = 4.0$, when the warranty period W is 2.0, the expected warranty costs are 2.09 cents per sales dollar, compared to 24.98 cents per sales dollar when the usage distribution is uniform. This is a reduction of about 92%.

One can analyze the effect of changes in the warranty period and the usage limit on warranty costs per unit sales. For example, for $m = 1.1$, for a usage limit of $U = 2.4$, if the warranty time period is doubled from $W = 2.0$ to 4.0, warranty costs per sales dollar increase from 1.55 to 1.80 cents, an increase of about 16%. On the contrary, for a warranty period of $W = 2.0$, if the usage limit is increased from $U = 2.4$ to 4.0, warranty costs per sales dollar increase from 1.55 to 2.28 cents, an increase of almost 47%. Management will have to weigh the different options and their benefits in order to select warranty policy parameters. For consumers that have high usage rates, they may prefer an increase in the usage limit in the warranty policy even at the expense of paying an added premium.

Fig. 3 shows graphs of expected warranty costs per unit sales, when the usage rate distribution is gamma with parameters $p = 4$, $m = 1.2$, and the

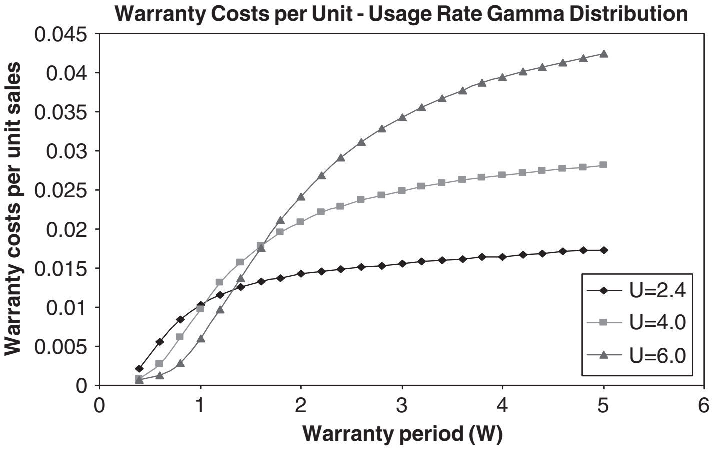


Fig. 3. Warranty Costs Per Unit Sales for Usage Rate having Gamma Distribution.

usage limit *U* being 2.4, 4.0, and 6.0, respectively, for different values of the warranty period *W*. For graphs of each value of *U*, warranty costs per unit sales increase with *W*, with the rate of increase becoming smaller for moderate to large values of *W*. For small values of *W* (say $W < 1.5$) an interesting phenomenon is observed. Unit warranty costs for larger values of the usage limit are less than those for smaller values of *U*. However, as the warranty period increases, the graphs for larger values of the usage limit dominate those for smaller values of *U*. This suggests that up to a certain cutoff value of the warranty period, management may offer a high usage limit as part of the warranty policy without a burden of increased unit warranty costs. This is a rather unique insight gained from the study.

A possible explanation of this phenomenon is as follows. Given that the usage rate has a gamma distribution with the selected values of the parameters, the joint effect of the value of the usage rate (which is a random variable) along with the two-dimensional warranty region, governed by the values of *W* and *U*, could lead to such results. Note that, for the chosen parameter values, the mean and standard deviation of the usage rate are 4 and 2, respectively. So, for small values of *W*, even if the value of *U* were to be large, say 6, the limit on the warranty time, *W*, could become the constraining factor if the usage rate was low (say 1). On the contrary, if the usage rate is quite high (say 8), the value of *U* could become the constraining parameter, which effectively reduces the limit on warranty time from the stipulated value of *W*. So, it is quite possible that the effective size of the two-dimensional region for a large value of U and a large value of the usage rate could be smaller than that with a smaller value of U and a smaller usage rate. This could lead to smaller values of warranty costs per unit for a situation where U is larger than another value, for the same value of W, when the warranty time period is small. As the value of W becomes larger, the probability that the value of U will become the constraining parameter in the two-dimensional region increases. This implies that the effective warranty time limit will be smaller than the value of W, when the usage rate is high. Furthermore, larger values of U will imply larger values of the warranty time limit in the two-dimensional region, influencing larger warranty costs per unit.

In many instances, in arriving at unit sales price for the product, the unit cost of production, unit warranty costs, and an expected profit margin are factors that come into play. Therefore, if a target profit margin is to be achieved, the decision maker may be forced to impose a bound on the unit warranty costs. Under such constraints, values of the parameters of a two-dimensional warranty policy may be determined using the results of this paper. Suppose that a bound of 1.5 cents per sales dollar has been imposed for expected warranty costs by management. From Fig. 3, observe that there are several combinations of warranty policy parameters that may satisfy this restriction. For example, a policy with a usage limit $U = 2.4$ and a warranty period W around 2.6, a policy with $U = 4.0$ and W around 1.2, and one with $U = 6.0$ and W around 1.4, would satisfy the warranty unit cost constraints. This analysis provides management with some options for choosing a policy, whereby additional criteria may influence the final selection of a suitable policy.

DISCUSSION

The results from the previous section may be used by management in several policy-related decision making areas. One particular area is the impact of quality improvement initiatives on warranty expenditures and thereby on the selection of warranty parameters. With an improvement in quality, the failure rate parameter will decrease. For the same warranty parameter values, this will lead to reduced warranty costs per unit sales. Management could estimate the total reduction in warranty expenditures and perform a cost/benefit analysis. An increased profit may result due to a larger reduction in warranty costs compared to the increase in quality improvement related expenses.

Such information may also be used in the context of product development to improve reliability of the product. Improved design will lead to an improvement in reliability and a reduction in the failure rate, eventually leading to a reduction in warranty expenditures. Cost figures will aid management in identifying the degree of the benefit of an improved design (improved profits due to reduced warranty costs versus increased product development costs) and hence provide a framework that indicates the value of such information. It could also be linked to compute figures on return on investment for such activities as product design and development, process improvement, and manpower training and development.

As a result of improved quality and better product design, product image in the marketplace improves. This could yield a larger market share for the company. With an increased sales figure, a larger profit could materialize. Alternatively, management may afford to increase the limits on the warranty policy parameters, thereby making the product stronger with respect to that of its competitors. The previous results have demonstrated the degree of tradeoff that could be made between the two warranty policy parameters for a given bound on the unit warranty costs.

Since the model incorporates a probability distribution for the usage rate, it provides the ability to identify expected warranty costs for different segments of users, thereby providing the motivation to offer a choice of warranty parameters to the user based on their habits. For a consumer with a high usage rate the manufacturer could actually provide a large value for U but a smaller value of W, which could be preferred by the consumer. On the contrary, for a consumer with a small usage rate, a large value of W could be preferred, which the manufacturer could oblige. Furthermore, if the usage rate is approximately constant, say as represented by a uniform distribution, the proportional difference between unit warranty costs for different values of U, for a given value of W, remains fairly constant for different values of W. The results also imply that once management has decided on a bound on the expected unit warranty costs, it could limit their choice on the value of the usage parameter, U. However, management might still be flexible on the selection of the warranty time parameter, W, as its impact is less dominant. As discussed in the previous section, management could use this information to select larger values of W, relative to that of its competitors, since the increase in unit warranty costs is only very moderate.

Estimating the usage rate distribution from historical data is of much value as the impact on expected warranty costs, as demonstrated previously, is quite significant. For consumers whose usage rate is characterized by a gamma distribution with a certain set of parameters, the unit warranty costs were found to be much lower than those having a uniform distribution of the usage rate. An adequate estimation of the usage rate distribution will aid the manufacturer in accurately estimating expected warranty costs and thereby expected profits. This approach enhances the developed model to incorporate an additional flexibility for market segmentation based on geographical regions or any other variable that is deemed to affect usage rate. Different usage rate distributions may be used for different market segments to arrive at an estimate of the unit warranty costs.

CONCLUSIONS

The paper has considered a two-dimensional warranty policy and derived expressions to estimate the expected warranty costs per unit sales. It is assumed that failed items within the region of warranty coverage are minimally repaired. Such an assumption is realistic for complex and expensive products since repair normally involves only a small part of the product. A good example in this setting is the automobile. On failure, the manufacturer has to repair the product (which is usually some parts and components of the automobile) and absorb repair costs. Warranty is usually not renewed on repair of the product. It is assumed, in minimal-repair, that the failure rate of the product remains unchanged after each repair.

Since warranty costs occur at a future point in time they have to be budgeted. Normally, a warranty reserves account is set up from which expenditures are disbursed. An accurate estimate of expected warranty costs is therefore necessary to ensure that the fund balance does not become negative. Given the availability of sales data, the paper has provided expressions for the expected warranty costs per unit sales.

Distributions of the usage rate have been assumed in the applications demonstrated in the paper. In practice, historical data from consumers would be used to validate an assumed distribution. Goodness-of-fit statistical tests may be used for such validation purposes. A similar comment applies to the choice of the failure rate function.

The paper has provided the decision-maker a framework to select warranty parameters for a two-dimensional policy, which is an important contribution. It provides insights on the tradeoffs between the two policy attributes, time and usage limit, for example. The financial implications on the choice of the two attributes are a major benefit of the study. Many warranty policies in practice are set based on the offerings of the competitors. This study provides a sound basis to aid in the selection of warranty parameters in a two-dimensional case.

REFERENCES

Ahn, C. W., Chae, K. C., & Clark, G. M. (1998). Estimating parameters of the power law process with two measures of failure rate. *Journal of Quality Technology*, *30*, 127–132.

Blischke, W. R., & Murthy, D. N. P. (1994). *Warranty cost analysis*. New York: Marcel Dekker.

Blischke, W. R., & Murthy, D. N. P. (1996). *Product warranty handbook*. New York: Marcel Dekker.

Chun, Y. H., & Tang, K. (1999). Cost analysis of two attribute warranty policies based on product usage rate. *IEEE Transactions on Engineering Management*, *43*(2), 201–209.

Eliashberg, J., Singpurwalla, N. D., & Wilson, S. P. (1997). Calculating the warranty reserve for time and usage indexed warranty. *Management Science*, *43*(7), 966–975.

Gertsbakh, I. B., & Kordonsky, K. B. (1998). Parallel time scales and two-dimensional manufacturer and individual customer warranties. *IIE Transactions*, *30*, 1181–1189.

Hsiung, C., & Shen, S. H. (2001). The effects of the warranty cost on the imperfect EMQ model with general discrete shift distribution. *Production Planning and Control*, *12*(6), 621–628.

International Mathematical and Statistical Library (IMSL), Inc. (1987). *Math/library user's manual*, *Version 1.0*. Houston, TX: IMSL.

Kim, H. G., & Rao, B. M. (2000). Expected warranty cost of two-attribute free replacement warranties based on a bivariate exponential distribution. *Computers and Industrial Engineering*, *38*, 425–434.

Lam, Y., & Lam, P. K. W. (2001). An extended warranty policy with options open to the consumers. *European Journal of Operational Research*, *131*, 514–529.

Moskowitz, H., & Chun, Y. H. (1988). *A Bayesian approach to the two-attribute warranty policy*. Paper No. 950. Krannert Graduate School of Management, Purdue University, West Lafayette, IN.

Moskowitz, H., & Chun, Y. H. (1994). A Poisson regression model for two-attribute warranty policies. *Naval Research Logistics*, *41*, 355–376.

Murthy, D. N. P., Iskander, B. P., & Wilson, R. J. (1995). Two dimensional failure free warranty policies: Two dimensional point process models. *Operations Research*, *43*, 356–366.

Singpurwalla, N.D. (1987). *A strategy for setting optimal warranties*. Report GWU/IRRA/Serial TR 87/4, The George Washington University, Washington, DC.

Singpurwalla, N. D., & Wilson, S. (1998). Failure models indexed by two scales. *Advances in Applied Probability*, *30*, 1058–1072.

Taguchi, G. (1986). *Introduction to quality engineering: Designing quality into products and processes*. Tokyo: Asian Productivity Organization.

Taguchi, G., Elsayed, E. A., & Hsiang, T. (1989). *Quality engineering in production systems*. New York: McGraw-Hill.

U.S. Federal Trade Commission Improvement Act, (1975). 88 Stat 2183. pp. 101–112.

Yeh, R. H., & Lo, H. C. (2001). Optimal preventive maintenance warranty policy for repairable products. *European Journal of Operational Research*, *134*, 59–69.

This page intentionally left blank

A COST VARIANCE INVESTIGATION USING BELIEF FUNCTIONS

Thomas W. Lin, Daniel E. O'Leary and Hai Lu

ABSTRACT

Using belief functions, this paper develops a model of the situation of a management team trying to decide if a cost process is in control, or out of control and, thus, in need of investigation. Belief functions allow accounting for uncertainty and information about the cost processes, extending traditional probability theory approaches. The purpose of this paper is to build and investigate the ramifications of that model. In addition, an example is used to illustrate the process.

1. INTRODUCTION

A production system, such as a manufacturing system, periodically is investigated to determine if the process is in control or out of control. Decision-making information typically is gathered from production and accounting information systems to determine the system status. Then that decision-making information is analyzed, often by multiple decision makers, and the control status of the system is determined, along with whether or

Applications of Management Science: In Productivity, Finance, and Operations
Applications of Management Science, Volume 12, 253–263

ISSN: 0276-8976/doi:10.1016/S0276-8976(06)12017-9

not the process is to be investigated. Unfortunately, determination of the control status is not a deterministic problem. As a result, there has been substantial analysis of the cost variance problem and a number of models have been proposed and developed to solve those cost variance problems. Researchers have employed probability models and fuzzy set models to investigate cost processes. The purpose of this research is to extend the cost variance model to a model-based on belief functions.

1.1. This Paper

This paper proceeds in the following manner. Section 2 reviews some of the previous literature associated with cost variance analysis models. Section 3 briefly reviews decision-making using belief functions. Section 4 reviews and summarizes some other applications of belief functions. Sections 5 and 6 discuss in detail an example designed to illustrate the use of belief functions in the cost variance problem. Finally, Section 7 summarizes the paper and reviews some potential extensions.

2. COST VARIANCE ANALYSIS

The basic cost variance investigation model is for a single period. In this model, it is assumed that the cost variance being controlled is either in control (ic) or out of control (oc). Management needs to determine if they should investigate a process based on a signal or a set of signals coming out of the process. This signal is typically set based on the production information and the accounting cost information. We also assume that there are multiple actors involved in the process, so that we need to aggregate their perspectives to generate a single overall strategy.

Lipe (1993) has discussed a number of issues of cost variance analysis, including framing of the problem and the effects of outcomes. Thus far, the primary mathematical models for cost variance analysis have been probabilistically based or fuzzy set based. Kaplan and Atkinson (1989) and Hirsch (1988) discuss the probability model in detail. The basic finding of the research is that cost variance processes have a critical value, where if the probability of the process being out of control exceeds that amount then it is cost beneficial to investigate the process. Zebda (1984) and Lin and O'Leary (1989) analyzed cost variance processes using fuzzy sets, extending the cost variance model from the classic probability-based analysis. Some of that research involved aggregating information from multiple actors to develop a single strategy. Oz and Reinstein (1995) contrast probability-based models and fuzzy set approaches.

There are some basic differences between probability, fuzzy set and belief functions and their application in the cost variance process. Using the probabilistic version, a process is either in control or out of control, with probabilities associated with state. With fuzzy sets, a process can be "almost" in control or "almost" out of control. We will see with belief functions that a process can be assessed as in control, out of control or it can be assessed that there is uncertainty as to whether the process is in control or out of control.

3. DECISION MAKING WITH BELIEF FUNCTIONS

The Dempster–Shafer theory of belief functions was formed in 1960s and 1970s (Shafer, 1976). Due to the difference between belief functions and probability framework in dealing with the uncertainties, belief functions framework are often regarded as more intuitive.

3.1. Background: Belief Functions

A frame Θ, is defined to be an exhaustive and mutually exclusive set of possible answers (a_i) to a particular question. If the question is a yes–no question, then there are only two answers, "yes" or "no."

Let $\Theta = (\{a_1, a_2, a_3\})$. Uncertainties are represented through m-values, which are equivalent to probabilities, but more general. The m-values are assigned to a subset of the elements of the frame.

For each Θ, let $m(\{a_i\})$, $m(\{a_i, a_j\})$, $m(\{a_i, a_j, a_k\})$, where $\Sigma_{A\subseteq\Theta}\, m(A) = 1$. That is, the m-values relate to each answer and interactions between answers. Define the belief function for A as Bel $(A) = \Sigma_{\,B\subseteq A}\, m(B)$. We set Bel $(\Theta) = 1$, and Bel $(\emptyset) = 0$.

The so-called "plausibility" for A is $\mathrm{Pl}(A) = 1 - \mathrm{Bel}\ ({\sim}A)$. Plausibility is the degree to which A is plausible in light of the evidence. It also can be thought of as the degree to which we do not disbelieve A.

3.2. Differences between Probability Theory and Belief Functions

The basic difference between m-values and probabilities is that m-values are assigned to subsets of the frame, whereas, probabilities are assigned to individual elements of the frame. Thus, probability theory and belief functions differ because probability is assigned to events, e.g., "in control" or "out of control." However, with belief functions, there also can be weight assigned as "uncommitted" to the process being either in control or out of control. Finally, Bel $(\{a\})+\text{Bel}\ (\{\sim a\})<1$, whereas in probability theory, $P(a)+P(\sim a)=1$, where $P(a)$ represents the probability of A.

3.3. Obtaining m Values

There are two basic ways to obtain m-values for a frame. First, the decision maker could directly assign values based on their judgment. Second, there may be some relationship between a frame where the weights have already been established and the current frame. In this last case, the existing information might be used to set the weights.

3.4. Dempster's Rule of Combination

The basic rule for combining independent items of evidence, while using belief functions, is Dempster's rule. Probability theory's "Bayes' rule" is similar to Dempster's rule, and is a special case of Dempster's rule. We will use this rule to combine belief functions, for example, across multiple decision makers.

Next, assume that we have multiple independent sets of Θ_j (e.g., multiple decision makers), where for each Θ_j the beliefs are denoted as m_j. Assume that the cardinality of Θ_j is 2. Let the combination of two pieces of evidence be denoted as m'. In that case,

$$m'(A) = \left(\sum\{m_1(b_1)\ \ m_2(b_2)|B_1 \cap B_2 = A,\ \ A \neq \varnothing\}\right)/K$$

In the case of combining belief functions from two sources I and O, the function could look as follows:

$$m'(I) = [m_1(I)*m_2(I) + m_1(I)*m_2(I,O) + m_2(I)*m_1(I,O)]/K$$

K is a constant, for normalization, and

$$K = 1 - \sum\{\underline{m}_1(b_1)\ \ m_2(b_2)|B_1 \cap B_2 = \varnothing\}$$

K is a term that is used to capture the "conflict" between two items of evidence. If $K=1$ then there is no conflict. If $K=0$ then the two items of evidence are in conflict with each other and cannot be combined.

For the case of combination of more than two pieces of evidence see Shafer (1976).

3.5. Statistical Information and Belief Information

Belief information can be combined with statistical information. If there are two sets of evidence, statistical and belief, then one approach to combining the two is to use Dempster's rule, treating the statistical data as one frame and the belief data as another frame. Similarly, combining probability and belief information can be done using Dempster's rule. For a further discussion see Srivastiva (1996).

4. APPLICATIONS OF BELIEF FUNCTIONS

The purpose of this section is to briefly review some of the primary applications of belief functions in business.

4.1. Applications of Belief Functions to Business Decisions

There have been a number of applications to information systems. Xie and Phoha (2001) applied belief functions and Dempster's rule in order to cluster groups of web users, in order to facilitate development of profiles. Srivastava (1999–2000) used belief functions to develop a model of Web-Trust Assurance for electronic commerce. Srivastava and Datta (2002) investigated the use of belief functions for the process of mergers and acquisitions. Other applications are discussed in Srivastava and Liu (2003) and in Shenoy and Srivastava (2003).

4.2. Applications of Belief Functions to Auditing

Many of the applications of belief functions have been developed for auditing. As a result, there has been a focus on using belief functions to analyze a number of issues (Srivastava, 1993, 1995a, 1997), including the following:

- The auditor's judgment process
- The strength of audit evidence
- The structure of audit evidence
- The risk of fraud
- Determining which opinion to give.

Srivastava (1995b) investigated auditor behavior, through value judgments using belief functions. A belief function model is developed that uses empirical data and predicts observed behavior discussed in the behavioral literature associated with "framing decisions" (e.g., Tversky & Kahneman, 1986). Srivastava and Mock (2000) reviewed how belief functions relate to behavioral research, such as the audit judgment process.

Shafer and Srivastava (1990) and Srivastava and Shafer (1992) investigated the basic nature and structure of audit evidence. In particular, they were concerned with how to model uncertainties involved in audit evidence, and understanding the structure of audit evidence.

4.3. Applications of Belief Functions to Management Accounting

Unfortunately, there has not been the same level of investigation of applications of belief functions in management accounting or other areas, as there has been in auditing. This paper addresses the use of belief functions to one of the primary problems associated with management accounting over the years. However, it is possible to extrapolate about its use. Most applications based on probability theory could be extended to use with belief functions. Further, Dempster's rule would facilitate combination of belief estimates across teams of users when a single estimate is necessary for decision-making.

5. MICROPROCESSOR CHIP PRODUCTION PROCESS

Sections 5 and 6 illustrate cost variance analysis using belief functions. The case study shows how to use belief functions from multiple decision makers in a semiconductor manufacturer to aggregate their assessments to derive a single estimate as to the belief that the process is in control.

Irvine Semiconductor Company is a firm that produces a super speed nanosecond microprocessor. The product goes through the following four-stage manufacturing process:

1. Fabricating

   The fabricating stage has three steps. First, raw materials, i.e., blank wafers are insulated with an oxide film. Second, the wafers are coated with a soft, light-sensitive plastic called photoresist. Third, wafers are masked by a glass plate (called a reticle) and flooded with ultraviolet light to pattern each layer on the water. These three steps are repeated to builds the necessary layers on the wafer. An array of transistors made up of various connected layers is called integrated circuits or chips.

The management accountant collects the wafer material and fabrication processing costs and the water fabrication yield rate information, and calculates the unit cost per wafer production.

2. Probing

   In the probing stage, an electrical performance test of integrated circuits is performed. The functioning chips are moved to the assembly stage.

   The management accountant collects the probing production cost and the probing yield rate information, and calculates both probe production cost per wafer and per chip.

3. Assembling

   The assembling stage has four steps. First, each wafer is cut into chips with a diamond saw. Second, the good chips are placed in a cavity of a ceramic package. Third, the bonding pads from the chips are connected by thin aluminum wires into leads of the package. Fourth, the package is sealed, with a metal lid placed over the exposed chips in the package.

   The management accountant collects the assembly product cost and the assembly yield rate information, and calculates the unit cost per chip.

4. Testing

   The packaged semiconductor device is tested to ensure that all electrical specifications of the integrated circuits are met.

   The management accountant collects the testing cost, test yield rate, and packaging cost information, and calculates the unit product cost per chip.

6. BELIEF FUNCTIONS AND COST VARIANCE INVESTIGATIONS

The company set the standard cost for each component of the microprocessor product with the monthly production of 200,000 units. During the first month, the company actually produced 190, 000 units. The accounting department provided the following unit chip cost information in Table 1.

Table 1. Cost Information.

Department	Standard Cost	Actual Cost	Variance	% Variance	%Variation
Wafer cost	\$40.00	\$42.00	\$2.00	2/32 = 6.25	5
Fabrication cost	150.00	162.00	12.00	12/32 = 37.5	8
Probe cost	80.00	93.00	13.00	13/32 = 40.625	16.25
Assembly cost	10.00	13.00	3.00	3/32 = 9.375	30
Test cost	30.00	31.00	1.00	1/32 = 3.125	3.33
Package cost	40.00	41.00	1.00	1/32 = 3.125	2.5
Total cost	\$350.00	\$382.00	\$32.00	100	100

Since the variance in the total cost is about 10% (32/350), management feels that some of the process variations need to be investigated. A three-person team is formed in order to proceed with the investigation. The team consists of a cost accountant, a supervisor, and a process engineer from each of the three subdivisions. They need to decide which processes should be investigated further and make proposal to the Vice President of Operations.

The team members agree on that they should focus on three processes: fabricating, probing, and assembling. The costs in these processes account for the majority of the total variation. For those three departments, the standard costs are \$230, and the variance is \$28. As a result, those three departments account for 87.5% of the total variance and the variance of those three is over 12%, compared to 3.6% for the rest of the processes.

However, the opinions of three members vary on which process is most likely out of control. Since no single opinion is dominant, they turn to a decision aid based on the belief functions to aggregate their assessments.

We assume that we are able to gather the *m*-values for each of the actors (supervisor, process engineer, and cost accountant) directly or we can assign *m*-values based on statistical evidence. We assume that there are three *m*-values, where the *m*-values are defined as

m(ic) – belief that the process is in control,
m(oc) – belief that the process is out of control, and
m(ic, oc) – uncertainty as to whether the process is in control or out of control.

Assume that the nine different sets of *m*-values, for each of the three actors are given in Table 2.

Table 2. m-values.

Department	Supervisor	Process Engineer	Cost Accountant
Fabrication process	m(ic) = 0.9, m(oc) = 0, m(ic,oc) = 0.1	m(ic) = 0.8, m(oc) = 0 m(ic,oc) = 0.2	m(ic) = 0.7, m(oc) = 0 m(ic,oc) = 0.3
Probe process	m(ic) = 0.8, m(oc) = 0, m(ic,oc) = 0.2	m(ic) = 0.1, m(oc) = 0.7, m(ic,oc) = 0.2	m(ic) = 0, m(oc) = 0.6, m(ic,oc) = 0.4
Assembly process	***m*(ic) = 0.8, *m*(oc) = 0, *m*(ic,oc) = 0.2**	m(ic) = 0.5, m(oc) = 0.4, m(ic,oc) = 0.1	m(ic) = 0.3, m(oc) = 0.2, m(ic,oc) = 0.5

We can see that for the Fabrication process, the belief that the process is out of control is 0 for all three, however, we can also see that that there is some uncertainty for each of them that the process is either in control or out of control. We will let $m(\text{ic})+m(\text{oc})+m(\text{ic, oc}) \ = 1$ in this case.

To illustrate, we make the calculations simple and adopt two-step approach in aggregating the assessments of three persons, a pair at a time. The first assessment is combined with the second assessment using the following formula.

$$m'(\text{ic}) = [m_1(\text{ic})*m_2(\text{ic})+m_1(\text{ic})*m_2(\text{ic, oc})+m_2(\text{ic})*m_1(\text{ic, oc})]/K$$
$$m'(\text{ic}) = [m_1(\text{oc})*m_2(\text{oc})+m_1(\text{oc})*m_2(\text{ic, oc})+m_2(\text{oc})*m_1(\text{ic, oc})]/K$$
$$m'(\text{ic, oc}) = [m_1(\text{ic, oc}) * m_2(\text{ic, oc})]$$

The combined assessment is then aggregated with the third assessment to obtain the total belief. The constant K is 1 when there is no conflict evidence in each assessment, and less than 1 when there is conflict evidence in each assessment. We see in fabrication process, all three persons assign 0 value to m(oc) though the belief in in-control is different. There is no obvious evidence showing that the process is out of control.

However, in probe and assembly processes, the assessments among three persons are significantly different.

For the Fabrication process, using Dempster's rule, we first combine the supervisor's and the process engineer's evidence to obtain $m'(\cdot)$.

$$m'(\text{ic}) = 0.9 *0.8+.9 * 0.2+0.8 *0\ .1 = 0.98$$
$$m'(\text{oc}) = 0 * 0+0 * 0.2+0 * 0.1 = 0$$
$$m'(\text{ic,oc}) = .01 * 0.2 = 0.02$$

Then using $m'(\cdot)$, and Dempster's rule, we combine the resulting combined supervisor's and process engineer's evidence with the cost accountant's beliefs to obtain the overall aggregated beliefs.

$$m''(\mathrm{ic}) = 0.98 \ast 0.7 + 0.98 \ast 0.3 + 0.7 \ast 0.02 = 0.994$$
$$m''(\mathrm{oc}) = 0 \ast 0 + 0 \ast 0.3 + 0 \ast 0.2 = 0$$
$$m''(\mathrm{ic, oc}) = 0.02 \ast 0.3 = 0.006$$

As a result, after combining the evidence across each of the supervisor, process engineer and cost accountant, the aggregate results are as follows:

For Fabrication process:	$m''(\mathrm{ic}) = 0.994,$	$m''(\mathrm{oc}) = 0,$	$m''(\mathrm{ic, oc}) = 0.006$
Probe process:	$m''(\mathrm{ic}) = 0.365,$	$m''(\mathrm{oc}) = 0.579,$	$m''(\mathrm{ic, oc}) = 0.056$
Assembly process:	$m''(\mathrm{ic}) = 0.878$	$m''(\mathrm{oc}) = 0.105,$	$m''(\mathrm{ic, oc}) = 0.017$

Assume that we investigate the process if the probability that the process is in control does not exceed 0.9. In that case, the recommendation is to investigate both probe and assembly process, but not the fabrication process since for the fabrication process, $m''(\mathrm{ic}) > 0.9$.

7. SUMMARY AND CONCLUSIONS

Belief functions are a powerful technique to model uncertain beliefs faced by managers and management accountants. Using belief functions we can aggregate beliefs as to whether a process is in control or out of control across multiple decision makers. Belief functions can be gathered directly from the participants or we could generate the values using other processes, such as statistical analysis.

REFERENCES

Hirsch, M. L., Jr. (1988). *Advanced management accounting*. Boston: PWS-Kent Publishing Company.

Kaplan, R. S., & Atkinson, A. (1989). *Advanced management accounting*. Englewood Cliffs, NJ: Prentice-Hall.

Lin, T. W., & O'Leary, D. E. (1989). Cost variance investigation with fuzzy sets. In: G. W. Evans, W. Karwowski & M. R. Wilhelm (Eds), *Applications of fuzzy set methodologies in industrial engineering* (pp. 213–222). New York: Elsevier.

Lipe, M. G. (1993). Analyzing the variance investigation decision: The effects of outcomes, mental accounting, and framing. *The Accounting Review*, *68*(October), 748–764.

Oz, E., & Reinstein, A. (1995). Investigation of cost variances: A contrast between the Markovian and fuzzy set approaches to accounting decision making. In: P. H. Siegel, De Korvin & K. Omer (Eds), *Applications of fuzzy sets and the theory of evidence to accounting*. Greenwich, CT: JAI Press Inc.

Shafer, G. (1976). *A mathematical theory of evidence*. New Haven: Princeton University Press.

Shafer, G., & Srivastava, R. (1990). The Bayesian and belief-function formalisms: A general perspective for auditing. *Auditing: A Journal of Practice & Theory*, *9*(Supplement), 110–137.

Shenoy, P. P., & Srivastava, R. P. (2003). Application of uncertain reasoning to business decisions: An introduction. *Information Systems Frontiers*, *5*(4), 343–344.

Srivastava, R. P. (1993). Belief functions and audit decisions. *Auditors Report*, *17*(Fall), 8–12.

Srivastava, R. P. (1995a). Evidential reasoning in auditing. Unpublished presentation at the AI/ES Research Workshop, Orlando, Florida.

Srivastiva, R. P. (1995b). Value judgments using belief functions. *Research on Accounting Ethics*, *2*(May), 109–130.

Srivastiva, R. P. (1996). Integrating statistical and non-statistical evidence using belief functions. In: A. Kent, J. G. Williams & C. M. Hall (Eds), *Encyclopedia of computer science and technology* (Vol. 37, pp. 157–174). New York: Marcel Dekker.

Srivastava, R. P. (1997). Audit decisions using belief functions: A review. *Control and Cybernetics*, *27*(2), 135–160.

Srivastava, R. P., & Datta, D. (2002). Belief-function approach to evidential reasoning for acquisition and merger decisions. In: R. P. Srivastava & T. Mock (Eds), *Belief functions in business decisions* (pp. 220–248). Heidelberg: Springer.

Srivastava, R.P., & Liu, L. (2003). Applications of belief functions in business decisions: A review. *Information Systems Frontiers*, *5*(4), 359–378.

Srivastava, R. P. (1999–2000). Evidential reasoning for web trust assurance services. *Journal of Management Information Systems*, *16*(3), 11–32.

Srivastava, R. P., & Mock, T. J. (2000). Belief functions in accounting behavioral research. *Advances in Accounting Behavioral Research*, *3*, 225–242.

Srivastava, R. P., & Shafer, G. R. (1992). Belief-function formulas for audit risk. *The Accounting Review*, *67*(April), 249–283.

Tversky, A., & Kahneman, D. (1986) Rational choice and the framing of decisions. *The Journal of Business*, *59*(4), Part 2, S251–S278.

Xie, Y., & Phoha, V. (2001). Web clustering from access logs using belief functions. *Proceedings of the 1st international conference on knowledge capture (K-CAP 01)*, ACM Press, 202–208.

Zebda, A. (1984). The investigation of cost variances: A fuzzy set theory approach. *Decision Sciences*, *15*, 359–388.

This page intentionally left blank

CHANCE-CONSTRAINED PROGRAMMING WITH SKEWED DISTRIBUTIONS OF MATRIX COEFFICIENTS AND APPLICATIONS TO ENVIRONMENTAL REGULATORY ACTIVITIES

William W. Cooper, Vedran Lelas and David W. Sullivan

ABSTRACT

This paper provides a theoretical framework for application of Chance-Constrained Programming (CCP) in situations where the coefficient matrix is random and its elements are not normally distributed. Much of the CCP literature proceeds to derive deterministic equivalent in computationally implementable form on the assumption of "normality". However, in many applications, such as air pollution control, right skewed distributions are more likely to occur. Two types of models are considered in this paper. One assumes an exponential distribution of matrix coefficients,

Applications of Management Science: In Productivity, Finance, and Operations
Applications of Management Science, Volume 12, 265–283

ISSN: 0276-8976/doi:10.1016/S0276-8976(06)12018-0

and another one uses an empirical approach. In case of exponential distributions, it is possible to derive exact "deterministic" equivalent to the chance-constrained program. Each row of the coefficient matrix is assumed to consist of independent, exponentially distributed random variables and a simple example illustrates the complexities associated with finding a numerical solution to the associated deterministic equivalent. In our empirical approach, on the other hand, simulated data typically encountered in air pollution control are provided, and the data-driven (empirical) solution to the implicit form of deterministic equivalent is obtained. Post-optimality analyses on model results are performed and risk implications of these decisions are discussed. Conclusions are drawn and directions for future research are indicated.

INTRODUCTION

This paper derives both explicit and implicit deterministic equivalents of a chance-constrained program in situations where matrix coefficients have skewed distributions. Prior uses of CCP (Charnes & Cooper, 1959; Charnes, Cooper, & Symonds, 1958; Charnes & Cooper, 1962; Charnes & Cooper, 1963; Allen, Braswell, & Rao, 1972; Jagannathan, 1985), have been confined mainly to normal distributions with randomness generally assigned to cost coefficients in the "objective" or right-hand sides. Experience in several applications, however, suggests that consideration needs to be given to skewed distributions and to randomness in the entire coefficient matrix. This is the direction taken in the present paper. Much of the past literature using CCP has assumed that the class of "zero-order" decision rules is the one that is applicable and we here continue in this tradition.

The paper is organized as follows. In the first section, we provide a brief overview of mathematical programming applications in air pollution modeling, with an emphasis on the problem of optimally determining pollution emission reduction rates (decision variables) for a number of pollution sources. The typical goal is to minimize the expected cost associated with these reductions and to satisfy government imposed pollution standards as chance constraints with prescribed risks of violation. Given the conditions likely to be encountered we want to meet these standards with high probability. This means that the standards are to be minimally met with very small chances of violation.

In the second section, we describe how the flow of toxic chemicals in air pollution management is often modeled using transfer coefficients between pollution sources and receptors. The pollution level at a receptor is generated by multiplying the transfer coefficient for that source-receptor pair by the emission level at the source. Source-receptor transfer coefficients can vary greatly because of variations in meteorological conditions such as wind direction and speed.

In the third section, instead of assuming a particular distribution for the transfer coefficients, we proceed by working directly with the data and deriving an empirical distribution function at every stage of the iterative process aimed at finding the optimal solution. Solutions and post-optimality analysis show that for skewed distributions of transfer coefficients, the solution obtained using the empirical approach satisfies the constraints with the desired probability. The last section concludes the paper with an outline of possible future research directions.

AIR POLLUTION MODELING

In this and the following section, we first outline the basics of air pollution meteorology and some of the most frequently used air quality modeling techniques. EPA (1980) defines an air quality model as a set of mathematical equations relating the release of air pollutants to the corresponding concentrations of pollutants in the ambient atmosphere. These models are used to predict the future concentrations of air pollutants and study their possible consequences. In general, air quality models are designed to address issues such as:

(i) Development of air pollution control plans for attainment and maintenance of acceptable air quality, mostly in urban areas where high pollutant concentrations are measured in residential neighborhoods near industrial zones with large pollution sources. Models are used to identify the sources contributing to the excessive concentrations. When such excessive concentrations are detected, air quality engineers are sent out to the plants to assess the situation and propose appropriate solutions. State agencies, such as Texas Commission on Environmental Quality (TCEQ), formerly Texas Natural Conservation Commission (TNRCC), then incorporate these solutions in the state air pollution control plan.

(ii) Assessment of expected environmental impacts from industrial expansion and urban development based on permit applications submitted to the state agency for approval, TNRCC (1994). Modeling is used by the state agency to predict how emissions from the new facility will affect ambient air quality.

(iii) Projections of the future air quality trends and patterns associated with regional planning operations in the rural areas of the state for several different options of industrial expansion. Air quality models can be used together with the economic analyses to rank each option as part of a planning process.

Mathematical programming models are extensively used in the air pollution management literature. See Greenberg (1995) and Cooper et al. (1996) for a survey and evaluation of these and other models. Generally the objective of these models is to optimize the cost of policy decisions with the great majority oriented to minimizing the cost of control and removal methods (Kohn, 1971; Kohn & Burlingame, 1971; Atkinson & Lewis, 1974; Ellis, McBean, & Farquhar, 1986). Usually, removal cost per ton of a pollutant is assumed to be given for a particular type of industrial plant like grain mills, chemical plants, petroleum refineries or power plants, Atkinson and Lewis (1974). Also, removal technologies are sometimes specified together with the costs involved in implementation and maintenance, and represented in the objective. Some models expand the objective to include other items such as the cost of fuel and its delivery, Morrison and Rubin (1985). Still other modeling approaches include damage functions in terms of human exposure to toxic chemicals that need to be balanced with economic consequences of necessary pollution reductions. Benefit functions like the latter are much more difficult to define and implement, so the cost of pollution removal continues to represent as a first choice for most modelers.

Over the years, governments at federal and state levels developed extensive lists of pollution standards. These standards can be classified roughly into "hard" (or no tolerance) standards that must be met without any exceedances, "moderate" standards that state how many times per period a standard can be violated, and "soft" (or "screening level") standards that are simply target levels for unregulated pollutants but which further additional examination into sources and health effects. The first type of standard naturally leads to deterministic constraints, while the other two suggest probabilistic constraints. Models dealing with SO_2 removal, Anadalingam (1987), Ellis, McBean, and Farquhar (1986), specify a wet sulfate deposition rate that should not be exceeded at a particular receptor. Under currently enforced regulatory standards, lead (Pb) faces a hard standard in that a calendar quarter average concentration cannot exceed a fixed value. Ground-level ozone and particulate matter are governed by moderate standards that allow multiple exceedances per year, which are combined over rolling three-year periods. Several volatile organic compounds (VOCs), like benzene and 1,3-butadiene, and other air toxics, such as arsenic, have specified effects screening levels (ESLs) concentrations for which exceedances are noted and forwarded to local authorities for their planning purposes, to enforcement staff who compare measured pollution levels to those expected under existing permits, and to monitoring operations personnel who may initiate further monitoring to look for higher levels nearby or to determine pollution sources. In David Sullivan (1997) and Vedran Lelas (1998), it was suggested that these pollutants be modeled with targets for the percent of days on which exceedances could be recorded.

One other important part of the literature is concerned with designs of optimal air pollution monitoring networks. Our models assume that a network of pollution monitors is given. There are several receptor selection procedures that have been developed in the literature. Seinfeld (1972), for example, describes the theory and computational algorithms that provide an optimal location for a given number of monitoring stations. The main criterion is that locations be chosen according to the sensitivity of concentration measurements to perturbations in source emissions. Langstaff, Seigneur, and Liu (1987) utilizes an air quality simulation model to produce representative air quality patterns, which is combined with population figures to obtain typical exposure patterns. These combined patterns are used to identify the most favorable locations for monitoring stations. A measure of a "sphere of influence" for proposed monitoring sites is also developed which makes it possible to eliminate redundant monitoring sites. An optimization procedure that determines the minimum number of sites required to achieve desired coverage. In Trujillo-Ventura and Ellis (1991) this development is taken as a step further by considering multiple objective functions that provide performance measures for spatial coverage of a network, together with abilities to detect violations of standards for multiple pollutants with associated costs and budget constraints.

To conclude the review of air pollution modeling, we point out that cost reduction, air quality standards, and transportation models based on diffusion processes are three basic elements of every air pollution model that uses management science techniques for decision making. In the next section, we discuss the later in some detail.

TRANSFER COEFFICIENTS

Our study focuses on the issue of examining benzene concentrations at particularly sensitive locations near large petrochemical plants and oil refineries in the Houston Ship Channel area. In order to reasonably approximate the concentrations at seven appropriately chosen monitoring stations (also referred to as receptors), the first task is to model the transfer coefficients, which represent pollution movement from source to receptor, for our sample of 44 largest benzene sources. We call the results of transfer coefficient simulation our basic or raw data and proceed to apply CCP model. The objective is to provide the state agency (TCEQ) with recommendations in terms of the cheapest optimal reduction rates for the sources that satisfy benzene health ESLs most of the time.

To generate transfer coefficients between sources and receptors, we use the *diffusion* (*dispersion*) *models*, where pollution concentrations (outputs) at a particular location are calculated using mathematical equations that depend on basic meteorological and geographical parameters (inputs) (Zannetti, 1990; Seinfeld, 1986). These models are based on the idealization of the physics of air transport and fluid flow. In Gaussian dispersion models the distribution of pollutants in the urban atmosphere is obtained by solving the following partial differential equation under appropriate initial and boundary conditions.

$$\frac{\partial C_j}{\partial t} + \nabla\left(uC_j\right) = -\nabla q_j + R_j + S_j \tag{1}$$

This equation, represents the mass balance of pollutant *j*, and must be verified in an elementary volume d*v* = d*x* d*y* d*z* in which the pollutant appears. Simply stated, this equation expresses a relation in which the accumulation of pollutant in volume d*v* during time d*t* is related to the pollutant inflow and outflow from this volume, and to the production and destruction rates of the pollutant during the same period, with $_{Cj}$ being the concentration of pollutant *j*, *u* the wind vector with components u_x, u_y, $u_{z,}$ q_j the mass flux of pollutant *j* due to turbulent diffusion, $_{Rj}$ the rate of production or destruction of pollutant *j* by chemical reaction, S_j the emission rate of pollutant *j*, ∇ the divergence operator ($\nabla = \partial u_x/\partial x + \partial u_y/\partial y + \partial u_z/\partial z$), and $\partial C_j/\partial t =$ partial derivative of the concentration with respect to time *t*.

The Gaussian model (1) is based on the statistical theory of wind fluctuations, and is constrained by the law of conservation of matter. For more details and complete derivations, see Guldman and Shefer (1980). We proceed to use this model while assuming a steady-state situation throughout a given time period. This steady-state situation is characterized by (i) constant source strength, (ii) constant wind speed and direction (we later relax this assumption), (iii) constant diffusion characteristics, and (iv) non-reacting pollutants. Despite these simplifying assumptions, this model is the one most used in air pollution, and its ability to capture impacts of principal meteorological parameters in a straightforward way contributes to its popularity. The model is expressed in a simple formula that describes concentration generated by point source under stationary meteorological and emissions conditions mentioned above. The pollution plume is usually set to be in direction of the *x*-axis, referred to as downwind direction, and the *y*-axis is referred to as crosswind direction. The specific version of the Gaussian plume formula used to generate our basic or raw data is as follows

$$C(s,r) = \frac{q}{2\pi\sigma_y\sigma_z\|u\|}\exp\left\{-\frac{1}{2}\left[\left(\frac{y_r}{\sigma_y}\right)^2 + \left(\frac{H}{\sigma_z}\right)^2\right]\right\} \tag{2}$$

In Eq. (2) $C(s,r)$ is the concentration at the receptor $r = \underline{(}x_r,y_r,0)$ due to emissions at the point source $s = (0,0,h_s)$; q is the emission rate at the source; σ_y and σ_z are standard deviations (horizontal and vertical) of the plume concentration distribution that is intended to capture the turbulence conditions as a function of distance d between source and receptor; $\bar{u}$is the average wind velocity vector; and $H = h_{s+}\Delta h$ is the effective emission height with Δh being the emission plume that rise above the source (see Fig. 1).

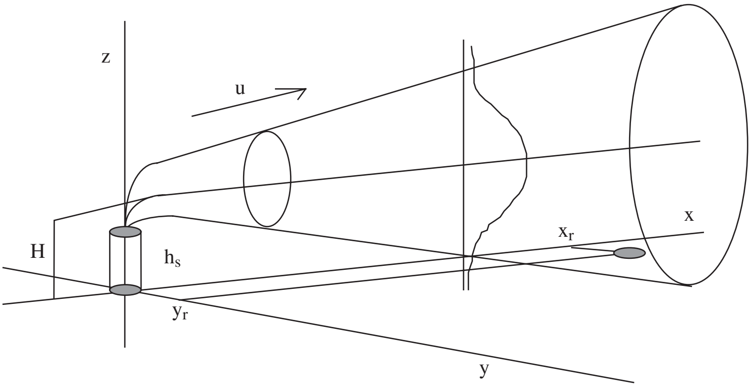


Fig. 1. Gaussian Plume Model.

From Eq. (2) and Fig. 1 we can see that all major meteorological factors (wind direction, wind speed, and atmospheric turbulence) as well as geographical factors (longitude, latitude, and stack height) are present. Since there is no time dependence, these meteorological variables must be considered homogeneous and stationary in the area between a source and receptor. Although an instantaneous plume distribution could be quite irregular, long average (one hour) concentrations in most cases exhibit bell-shaped Gaussian distributions in both horizontal and vertical directions.

The main concern when using the Gaussian model is accuracy in estimation of its parameters, in particular the diffusion coefficients σ_y and σ_z. There are two principal ways of estimating diffusion coefficients; the first method calculates these parameters using standard deviations of crosswind and vertical wind vector components, and nondimensional functions of distance or time traveled between the source and receptor; the second method uses tables to ascertain stability of the atmosphere using various meteorological parameters, and then, given the stability classification, calculates diffusion coefficients σ_y and σ_z. There are a number of extensions and special cases for which some variations of basic Eq. (2) is used, but they are not discussed here. See Zannetti (1990) for more details.

The way we use the Gaussian dispersion model will be described in more detail below, but the basic idea is to use historical wind direction and speed daily (or even hourly) data, calculate diffusion coefficients and evaluate Eq. (2) for a given source-receptor pair. Hence, using historical meteorological data, we obtain a distribution of concentrations for each source-receptor pair to obtain better understanding of the relationships between them. We then use these concentrations as an input to our CCP model in ways we describe next.

We now turn to description of the simulation experiment, which we used to obtain the data on transfer coefficients for our sample of sources and monitors. Emissions data for selected sources were taken from the set that the TCEQ maintains as part of its Point Source Data Base on large benzene emission sources in the Houston Ship Channel area. The data elements for the sources provide the latitude and longitude of each emission point, the type of source (usually a plant seen as a refinery) and estimated quantity of emission in grams per second. These emissions can vary considerably based on the activity level in a plant as well as exogenous factors such as temperature. In our simulation experiment, however, we will assume constant values, q, for these current emission levels. We use pollution sources with high benzene emissions that are located in a vicinity of the selected seven monitoring stations (Table 1).

Table 1. Houston Regional Monitoring Network Coordinates.

HRM Monitor Identifier	Longitude	Latitude
801	−95.256	29.714
803	−95.181	29.765
804	−95.132	29.835
807	−95.017	29.769
808	−95.056	29.646
815	−95.027	29.753
901	−94.910	29.766

Table 2. Benzene Sources in Houston Ship Channel Area.

Name	Longitude	Latitude	Benzene (g/s)	Name	Longitude	Latitude	Benzene (g/s)
S2	−95.4615	29.99405	0.179535	TOL905	−95.1275	29.71281	0.406561
S4	−95.4614	29.99400	0.172558	FUGACU	−95.1217	29.72234	0.249873
CD-1	−95.4614	29.99588	0.215944	F34E00	−95.1216	29.83246	0.648919
FE-14A	−95.2562	29.72514	0.341388	CUMFE	−95.1214	29.72606	0.325555
FU40ADUTKF	−95.2354	29.71347	0.383644	F44E00	−95.1206	29.82916	0.278076
FU93BT	−95.2348	29.71076	1.304756	F40E00	−95.1192	29.83250	0.383213
FU19DOCKC	−95.2346	29.72109	0.269619	_774E14	−95.1185	29.81661	0.175564
FU91ISOMI	−95.2338	29.71139	0.206443	F25E00	−95.1182	29.82799	1.076956
F-ARU	−95.2338	29.71365	0.843411	_38E01	−95.1182	29.82799	0.194032
FU62HDS	−95.2338	29.71034	0.269888	F8E00	−95.1131	29.83306	0.240769
F-733	−95.2330	29.71123	0.623874	_17E42	−95.1094	29.83309	0.428082
F-BTU	−95.2317	29.71054	1.829040	_718E05	−95.1035	29.81745	0.252572
F-12-BIO	−95.2313	29.72120	0.489916	_723E02	−95.1022	29.81889	0.300203
F-10-API	−95.2311	29.71960	1.493848	ZACET−7	−95.1015	29.72892	0.434107
FUG-WEST	−95.2043	29.73054	0.279363	ZACET−1B	−95.1015	29.72920	1.073059
FUG-CENTER	−95.1998	29.72854	0.187405	SD−1	−95.0778	29.76648	0.364076
MARINE	−95.1983	29.72844	0.218798	FGNRU	−95.0168	29.74662	0.210642
SVP−1	−95.1798	29.75741	0.234177	WOUFUG	−95.0141	29.72389	0.260020
FUGDISP	−95.1348	29.72023	0.463740	E−1−17	−95.0090	29.75447	0.162988
PY3FUG	−95.1314	29.71419	0.190760	DOCKSSHIP	−95.0068	29.74062	0.783232
HT2FUG	−95.1303	29.71441	1.759757	COUD40	−95.0040	29.74518	0.174404
CIPXFUG	−95.1276	29.71596	0.286384	F−1592−31	−94.9218	29.82508	0.316353

These 44 pollution sources (Table 2), consist mainly of oil refineries and petrochemical industries, most of which are located along the Houston Ship Channel area. Our set of monitoring stations is located downwind from most of the sources and residential neighborhoods so that the impact of benzene emissions can be properly assessed.

Transfer coefficients at the receptors (which represent transmissions from the sources) form an m (number of receptors) by n (number of sources) coefficient matrix for our CCP model. Each transfer coefficient $a_{ij(\omega)}$ represents a marginal increase in pollution concentration at receptor i for a unit increase in emissions at source j. For our simulation, these coefficients are generated by the Gaussian model (2), that takes into account possible variations due to elements such as distance from source to receptor, height of the emission source, wind speed and direction, and the diffusion coefficients that describe how a plume of pollution spreads from a point source, Guldman and Shefer (1980). Most of these elements exhibit considerable variation in course of a year and, hence, are reflected in random behavior for the transfer coefficients. We have examined a year's worth of daily transfer coefficients generated by Gaussian simulation for 2002 meteorological data from the TCEQ database. We found that the distributions are highly skewed to the right, which motivated the use of empirical distributions.

Distribution of a typical transfer coefficient has a mean of order 10^{-2} and is about one hundred times as large as the median. Some of the medians are indeed very close to zero. This is because a considerable probability mass of transfer coefficient distribution is at zero, indicating that for a significant period of time there was no measurable transfer from a source to a receptor. Larger means (and medians) are usually found in areas where sources cluster around downwind receptors such as 801, 804, and 815 (see Fig. 2 which displays the exact geographical position of seven monitors and 44 sources in our sample). Variance of a typical transfer coefficient can be very large, particularly in those cases where we have a bimodal distribution with the gap separating small and large readings.

On the other hand, smaller values are associated with more isolated monitoring stations such as 807, 808, and 901. Variation follows similar pattern. It is relatively large for receptors 801 and 804 and quite small for the others. Skewness is very pronounced at all monitoring stations mainly because of a small fraction of "moderate" days producing relatively large transfer coefficient readings and even smaller fraction of "critical" days producing very high observations most of which typically exceed benzene effect screening levels.

Examination of these preliminary observations on transfer coefficient distributions indicates their extremely irregular nature. Fig. 3 portrays this behavior for transfer coefficient between source identified as S2 and monitoring station 801.

Some of the distributions are quite close to a classical bimodal structure with one (larger) cluster of the assumed values at the lower end and another

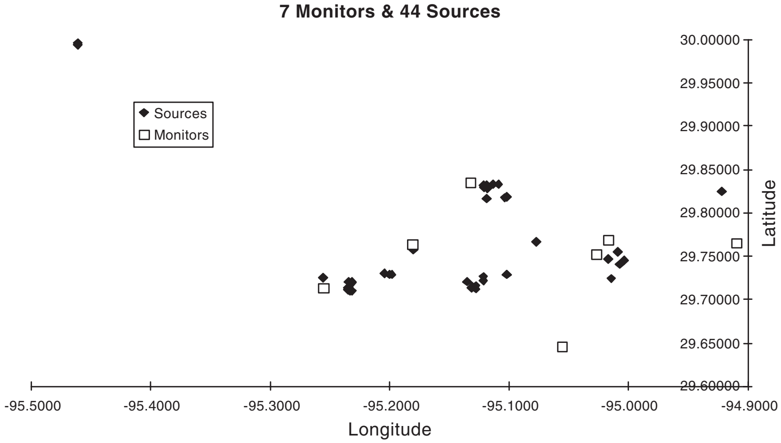


Fig. 2. Monitors and Sources in Houston Ship Channel Area.

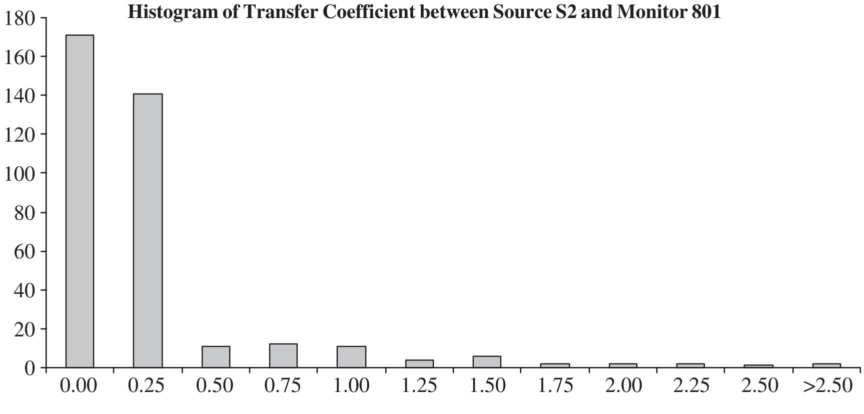


Fig. 3. Transfer Coefficient Distribution.

(smaller) cluster of observations at the higher end. The reason for this behavior lies in meteorological parameters that enter as inputs into transfer coefficient calculations. In the Houston Ship Channel area, winds blow from the south most of the time. However, the unpredictability of wind shifts (in direction as well as speed), play a major role in estimating transfer coefficients between the sources in Table 2 and the monitors in Table 1. Wind direction is the dominant meteorological condition to be considered, particularly on those days when receptors are directly downwind from a large number of sources. For such instances large transfer coefficient readings are more likely to occur when compared to the other days in a year.

The question then becomes how to summarize such distributions effectively so that most of the relevant information is preserved. The best approach is to use empirical cumulative distributions for each transfer coefficient, and utilize their percentiles in our CCP models. In our experiment, the goal is to find the emission reduction rates for each pollution source by minimizing the cost of these reductions and satisfying government imposed pollution standards with a prescribed high probability. For this purpose, we interpret our model (1) from the first section as follows. The decision variables x_j, for $j = 1,\ldots,n$ are reduction rates, and hence are restricted to be between zero and one. Given c_j, the per unit reduction cost for each source, our objective function represents the total of such reduction costs.

The constraints we will use are slightly different from those in model (1). In this modification the current emission levels q_j for source j are multiplied by 1-x_j, the after-reduction emission rates. This enables us to obtain the pertinent emission levels $q_j(1$-$x_j)$, which are then placed in the constraints instead of x_j. For complete model description, refer to the next section in this paper.

In the next section, we outline the empirical approach to solving CCP model (1), which proves particularly useful when dealing with the types of skewed distributions we encountered here.

CCP MODEL WITH EMPIRICAL DISTRIBUTION

We start with the general formulation of a CCP model in (1), and concentrate on $A(\omega)$ an m x n matrix of random variables with unknown distribution functions $F_{a_{ij}(\omega)}$. Our assumption is that we only have a certain amount of data on each of the matrix coefficients, and furthermore, we assume that, whatever the true underlying distributions are, they are not normal but heavily skewed to the right, which confirms with the empirical findings in air pollution monitoring described in the previous section.

If we knew each of the $F_{a_{ij}(\omega)}$ in closed form, we could write the deterministic equivalent of (1) as follows:

$$\begin{aligned} & opt \quad f(x) \\ s.t. \quad & F_{Y_i(\omega,x)}(b_i) \geq \alpha_i, i = 1, K, m \\ & l \leq x \leq u \end{aligned} \tag{3}$$

with $Y_i(\omega, x) = \Sigma_{j=1}^{n} a_{ij}(\omega)x_j$, and $F_{Y_i(\omega,x)}$ as left-hand side distribution functions, for $i = 1,\ldots,m$.

The problem is that even if we knew each of the $F_{a_{ij}(\omega)}$'s, we might still not be able to find the closed form of $F_{Y_i(\omega,x)}(b_i)$ as a function of a vector of decision variables x. Indeed, that turned out to be an extremely difficult task even for a simple right-skewed distribution such as exponential (see first two sections). The core of the problem is in assessing the distribution of a linear combination of random variables. If all the random variables are normal, then the linear combination of them is also normal. This property does not hold for the type of skewed distributions we are considering.

Furthermore, even if we are able to overcome this difficulty by means of approximations, we are still facing, in general, a nonconvex and probably nondifferentiable mathematical programming problem. Therefore, we can only formulate a general deterministic equivalent based on (3), emphasizing implicit nature of the constraint functions as follows

$$\begin{aligned} opt \quad & f(x) \\ s.t. \quad & g(x) \leq 0 \\ & l \leq x \leq u \end{aligned} \tag{4}$$

where $f(x)$ is the objective function, vectors l and u are lower and upper bounds on decision vector $\mathbf{x}$, and $g(x) = (g_1(x), K, g_m(x))$,with component functions defined as $g_i(x) = \alpha_i - F_{Y_i(\omega,x)}(b_i)$. These constraint functions are implicit in decision vector x and will be the main focus of our investigation.

We want to emphasize again that we do not know the constraint functions g_i explicitly. All we have is a data set on each of the matrix coefficients. We need to develop an iterative algorithm that uses empirical distribution of the ith linear constraint left-hand side $Y_i(\omega,x)$, at the particular iteration vector x^{k} to evaluate $g_i(x^k)$, and which, starting from some initial point x^0, converges to a relative optimum x^*.

Since we do not know the explicit form of the constrained functions, we cannot investigate their properties such as continuity, differentiability, and convexity. However, if carefully implemented, this empirical approach can provide us with the relative optimum for all kinds of distributions of our matrix coefficients. The most important implication is that it is possible to solve this kind of problem having only the data on matrix coefficients and without assuming any theoretical distributions. This means that we can have all kinds of skewed distributions, bimodal distributions, and whatever data turns out to be. As such, the risks of satisfying the constraints will be more precisely represented than would be the case when using theoretical distributions that do not fit the data well. Further, in our empirical approach, we do not move toward normality, as center limit theorem often does, and this enables us to truly concentrate on the tails of our distributions where the risks of constraint violations lie.

The most common approach to determine empirical distributions for a set of data is to use a piecewise linear function. This function is linear on the adjacent intervals and has the same ordinate at the interval endpoints. The drawbacks of this approach are the assumption of linearity, which can be partially circumvented using very fine grid of intervals at additional computational expense, and the fact that the first derivatives are discontinuous at the interval endpoints.

The following methodology is based on Section 6.2.4, in Law and Kelton (1991, p. 350). Suppose that we have N observations on each transfer coefficient $a_{ij}(\omega)$ denoted as $a_{ij}^1, a_{ij}^2, K, c_{ij}^N$. Since each of the transfer coefficients is a nonnegative random variable, and decision variables are assumed to be nonnegative, we can let the leftmost endpoint to be $s_0 = 0$. Since the values of constraint left-hand side $Y_i(\omega,x)$ depend on the current value of decision vector x, we define the flexible interval width $h_i(x) := M_i(x)/20$, where 20 is chosen to be the number of intervals. Here, $M_i(x)$ is the maximum over all N data points of the constraint left-hand side $Y_i(\omega,x)$ for a particular decision vector x. That is,

$$M_i(x) := \max_{1 \le k \le N} Y_i^k(x),\ Y_i^k(x) := \sum_{j=1}^{\mathrm{n}} a_{ij}^k x_j \tag{5}$$

Then we have $s_{it}(x) := tM_i(x)/20$, for $t = 0,1,\ldots,20$. In order to define the values of our empirical distribution function $F_{Y_i(\omega,x)}(s)$, $0 \le s \le M_i(x)$, of the left-hand side $Y_i(\omega,x)$ in each of the endpoints $s_{it}(x)$, $t = 0,1,\ldots,20$ we need to count the number of observations in each interval $(s_{i,t-1}(x), s_{it}(x))$, $t = 1,2,\ldots,20$ and then calculate the cumulative count. Let

$$N_{it}(x) := \left|\{k|\,Y_i^k(x) \ge s_{i,t-1}(x) \& Y_i^k(x) < s_{it}(x)\}\right| \tag{6}$$

represent the number of data points in t-th interval, $t = 1,2,\ldots,20$. Then the value of empirical CDF at the endpoint $s_{it}(x)$, $t = 0,1,\ldots,20$ is given by

$$F_i(s_{it}(x)) := F_{Y_i(\omega,x)}(s_{it}(x)) = \frac{\Sigma_{h=1}^t N_{ih}(x)}{N} \tag{7}$$

Using (7) we have the following expression for piecewise linear CDF:

$$F_i(x,s) = \left(F_i(s_{i,t-1}(x)) - s_{i,t-1}(x)\frac{F_i(s_{it}(x)) - F_i(s_{i,t-1}(x))}{s_{it}(x) - s_{i,t-1}(x)}\right) + \frac{F_i(s_{it}(x)) - F_i(s_{i,t-1}(x))}{s_{it}(x) - s_{i,t-1}(x)}s, s_{i,t-1}(x) \le s \le s_{it}(x) \quad (8)$$

Now, let

$$A_i(x) = F_i\left(s_{i,t-1}(x)\right) - s_{i,t-1}(x)\frac{F_i(s_{it}(x)) - F_i\left(s_{i,t-1}(x)\right)}{s_{it}(x) - s_{i,t-1}(x)} \quad (9)$$

be the intercept, and

$$B_i(x) = \frac{F_i(s_{it}(x)) - F_i\left(s_{i,t-1}(x)\right)}{s_{it}(x) - s_{i,t-1}(x)} \quad (10)$$

the slope of linear function on interval ($s_{i,t}$$-_1(x)$, $s_{it}(x)$) that contains b_i. Then the *i*th constraint in (3) is represented with

$$A_i(x) + B_i(x)b_i \ge \alpha_i \quad (11)$$

This leads to an empirical deterministic equivalent

$$\begin{aligned} & opt \quad f(x) \\ s.t. \quad & A_i(x) + B_i(x)b_i \ge \alpha_i, \quad i = 1,\ldots,m \\ & l \le x \le u \end{aligned} \quad (12)$$

Now, we provide a description of how we implemented model (12). As we mentioned before, we start with an initial decision vector x^0 that is as close to being feasible as possible. Because the feasible region is just a fraction of [0,1] hypercube, our intuition tells us that this initial vector should be "close" to the "true" optimal solution, the one where cost is minimized, and all the constraints are satisfied in with less than 5% of violations. Assuming we are at the *k*-th iteration of our algorithm, we need to calculate $A_i(x^k)$ and $B_\text{i}(x^k)$ using (9) and (10), maintain feasibility in (12) and move toward x^{k+1} where the cost is lower. In order to calculate intercept and slope in (9) and (10) respectively, we need to know the grid $s_{it}(x^k)$, $t = 0,1,\ldots,20$ and CDF values $\text{F}_i(s_{it}(x^k)))$. Grid calculation depends on the maximum $M_i(x^k)$, which is updated dynamically, and CDF values are calculated using (7) where the cumulative counts are determined once the grid is known.

The entire methodology was implemented in excel using the raw data on transfer coefficients and *max*, *sum*, *count*, and *if* functions built in the spreadsheet. The results heavily depended on starting vector but, starting with optimal the decision vector "reasonably" close to the feasible region, we were able to compare standard model (2) which assumes normally distributed coefficient matrix, with (12), the empirical deterministic equivalent of (1). Table 3 shows the initial vector, together with the solutions to (2) and (12).

As we can see from Table 3, the standard model results in much lower level of reductions, compared with both the initial vector, and especially with the results from the empirical model. The percentage violations for both models, as well as for the initial vector, are given in Table 4.

Although the empirical model clearly outperforms the standard model, especially at monitor station 801, where the percentage of violations is much higher than the required 5%, we can see that these violations are not uniformly distributed across monitoring stations. There are few stations (801

Table 3. Initial Vector and Optimal Solutions.

Name	Initial	Standard	Empirical	Name	Initial	Standard	Empirical
S2	0.0000	0.0000	0.0000	TOL905	0.0000	0.0000	0.0164
S4	0.0000	0.0000	0.0056	FUGACU	0.0000	0.0000	0.0544
CD-1	0.0000	0.0000	0.0082	F34E00	1.0000	0.8551	0.9806
FE-14A	1.0000	0.5281	0.9899	CUMFE	1.0000	0.3745	0.9905
FU40ADUTKF	1.0000	0.4133	0.9885	F44E00	0.0000	0.0000	0.3459
FU93BT	0.0000	0.0000	0.3611	F40E00	0.0000	0.0000	0.0016
FU19DOCKC	0.0000	0.0000	0.0541	774E14	0.0000	0.0000	0.0000
FU91ISOMI	0.0000	0.0000	0.0000	F25E00	0.0000	0.0000	0.1774
F-ARU	1.0000	0.9835	0.9748	38E01	0.0000	0.0000	0.0000
FU62HDS	1.0000	0.0000	0.9920	F8E00	1.0000	0.3197	0.9928
F-733	1.0000	0.0000	0.9904	17E42	0.0000	0.0000	0.0299
F-BTU	0.0000	0.0000	0.1503	718E05	0.0000	0.0000	0.0000
F-12-BIO	0.0000	0.0000	0.1286	723E02	1.0000	0.7520	0.9910
F-10-API	1.0000	0.5477	0.9553	ZACET-7	1.0000	0.2844	0.9871
FUG-WEST	0.0000	0.0000	0.0000	ZACET-1B	1.0000	0.8138	0.9679
FUG-CENTER	0.0000	0.0000	0.0356	SD-1	0.0000	0.0000	0.0077
MARINE	1.0000	0.8572	0.9935	FGNRU	0.0000	0.0000	0.0108
SVP-1	0.0000	0.0000	0.0000	WOUFUG	0.0000	0.0000	0.0000
FUGDISP	1.0000	1.0000	0.9861	E-1-17	0.0000	0.0000	0.0000
PY3FUG	0.0000	0.0000	0.0177	DOCKSSHIP	0.0000	0.0000	0.3216
HT2FUG	0.0000	0.0000	0.0080	COUD40	0.0000	0.0000	0.0000
CIPXFUG	0.0000	0.0000	0.0000	F-1592-31	0.0000	0.0000	0.2676

Table 4. Percentage Violations for Initial Vector and Optimal Solutions.

HRM Monitor Identifier	Initial (%)	Standard (%)	Empirical (%)
801	6.3	16.2	5.2
803	3.8	11.0	2.5
804	0.0	11.5	0.8
807	5.8	6.0	2.5
808	2.5	2.5	0.8
815	9.3	11.2	4.1
901	3.6	3.6	3.3

and 815) that have close to 5% violations, whereas some others (804 and 808) have less than 1%. This is partly due to the fact that monitoring stations 801, 804, and 815 are close to large cluster of sources. The puzzling lack of violation at 804 can be somewhat explained by high variability of transfer coefficient distributions. On the other hand, stations 807, 808, and 901 are relatively further away from any major cluster of sources, and hence experience fewer violations.

CONCLUSIONS

The first part of the paper presented the theory of CCP models with random and skewed coefficient matrix using independent exponentially distributions. The problem with Poisson, lognormal, chi-squared, and most other skewed distributions are that cumulative distribution function is not known in the closed form. Furthermore, their characteristic functions of these distributions are not as simple as the one for exponential distribution, preventing us from calculating necessary integrals when deriving the density of the left-hand side distributions. Further research in this area should concentrate on examining numerical properties of deterministic equivalent in (26), and investigate under which conditions can global optima be achieved.

The second part of the paper presented the empirical approach to solving the same CCP model, and then applied it to the air pollution control problem of determining optimal pollution reduction levels, subject to air quality standards being satisfied most of the time. We have demonstrated that this approach is clearly superior to the standard assumption of normality in terms of the percentage of violations being much closer to desired levels. Further research should consider revising deterministic equivalent in (12) with perhaps a cubic spline function, which would assure the smoothness in the empirical CDFs, as well as better representation of the first three moments of skewed distributions. Developing a more robust implementation of the deterministic equivalent, would also contribute to better quality of the solution.

REFERENCES

Charnes, A., & Cooper, W. W. (1959). Chance-constrained programming. *Management Science*, *6*(1), 73–79.

Charnes, A., Cooper, W. W., & Symonds, G. H. (1958). Cost horizons and certainty equivalents: An approach to stochastic programming of heating oil production. *Management Science*, *4*(3), 235–263.

Charnes, A., & Cooper, W. W. (1962). Chance-constraints and normal deviates. *Journal of American Statistical Association*, *57*(297), 134–148.

Charnes, A., & Cooper, W. W. (1963). Deterministic equivalents for optimizing and satisfiying under chance constraints. *Operations Research*, *12*(3), 460–470.

Allen, F. M., Braswell, R. N., & Rao, P. V. (1972). Distribution-free approximations for chance constraints. *Operations Research*, *22*, 610–621.

Jagannathan, R. (1985). Use of sample information in stochastic recourse and chance constrained programming. *Management Science*, *31*(1), 96–108.

TNRCC. (1994). Stationary source and CEMS test observation and test reporting review protocol, Austin, TX.

Zannetti, P. (1990). *Air pollution modeling*. New York: Van Nostrand Reinhold.

Seinfeld, J. H. (1986). *Atmospheric chemistry and physics in air pollution*. New York: Wiley.

Guldman, J. M., & Shefer, D. (1980). *Industrial location and air quality control*. New York: Wiley.

Greenberg, H. J. (1995). Mathematical programming models for environmental quality control. *Operations Research*, *43*(4), 578–622.

Cooper, W. W., Hemphill, M. W., Lelas, V., Li, S. X., Huang, Z. M., & Sullivan, D. W. (1996). Survey of mathematical programming models in air pollution management. *European Journal of Operational Research*, *96*, 1–35.

Kohn, R. E. (1971). Optimal air quality standards. *Econometrica*, *39*(6), 983–995.

Kohn, R. E., & Burlingame, D. E. (1971). Air quality control model combining data on morbidity and pollution abatement. *Decision Science*, *2*, 300–310.

Atkinson, S. E., & Lewis, D. H. (1974). A cost-effective analysis of alternative air quality control strategies. *Journal of Environmental Economics and Management*, *1*, 237–250.

Ellis, J. H., McBean, E. A., & Farquhar, G. J. (1986). Chance constrained/stochastic linear programming model for acid rain abatement – II. Limited colinearity. *Atmospheric Environment*, *20*(3), 501–511.

Morrison, M. B., & Rubin, E. S. (1985). A linear programming model for acid rain policy analysis. *Journal of Air Pollution Control Association*, *35*(1), 1137–1148.

Anadalingam, G. (1987). A multiple criteria decision analytic approach for evaluating acid rain policy choices. *European Journal of Operational Research*, *29*, 336–352.

Seinfeld, J. H. (1972). Optimal location of pollutant monitoring stations in an airshed. *Atmospheric Environment*, *6*, 847–858.

Langstaff, J., Seigneur and, C., & Liu, M. K. (1987). Design of an optimal air monitoring network for exposure assessment. *Atmospheric Environment*, *21*(6), 1393–1410.

Trujillo-Ventura, A., & Ellis, J. H. (1991). Nonlinear optimization of air pollution monitoring networks: algorithmic considerations and computational results. *Engineering Optimization*, *19*, 287–308.

Law, A. M., & Kelton, W. D. (1991). *Simulation modeling and analysis* (2nd ed). New York: McGraw-Hill.

Sullivan, D.W. (1997). Ph.D. Thesis, University of Texas, Austin, TX.

Vedran Lelas, (1998), Ph.D. Thesis, University of Texas, Austin, TX.

This page intentionally left blank

OPTIMAL ADVERTISING PULSATION POLICY FOR A CONTINUOUS MODEL OF ADVERTISING COMPETITION

Hani I. Mesak and Hongkai Zhang

ABSTRACT

Based on a continuous version of the Lanchester advertising model for a duopoly, a mathematical model is developed to determine the optimal advertising policy of a firm responding to the advertising pulsation policy of its competitor. A Dynamic Programming (DP) approach has been employed to arrive at the optimal solution.

It has been mainly demonstrated that under a concave or linear advertising response function, the focal firm's DP policy is superior to its Uniform Advertising Policy (UAP) counterpart (constant advertising spending over time), irrespective of the advertising pulsation policy employed by its rival. Under a convex advertising response function, on the other hand, the focal firm's DP policy is superior to its Advertising/Maintenance Pulsing Policy (APMP) and Advertising Pulsing Policy (APP) counterparts (alternating advertising spending at two levels), irrespective of the advertising pulsation policy used by the competitor.

Applications of Management Science: In Productivity, Finance, and Operations
Applications of Management Science, Volume 12, 285–308

ISSN: 0276-8976/doi:10.1016/S0276-8976(06)12019-2

INTRODUCTION

Advertising is a key factor in a firm's marketing efforts to which significant resources are usually committed. Thus, the determination of an optimal advertising policy with respect to a certain performance measure is of central importance to both professionals and academicians. An important concern of advertising managers is allocating an advertising budget effectively over time. Marketing researchers have analyzed a variety of models to address the issue of whether or not it is best to adopt a *pulsing* policy, alternating between low and high levels of spending of a finite frequency, or an *even* policy, scheduling advertising exposures at a constant rate (Feinberg, 1988; Hahn & Hyun, 1991; Luhmer, Steindl, Feichtinger, Hartl, & Sorger, 1988; Mahajan & Muller, 1986; Mesak, 1985; Mesak & Darrat, 1992; Park & Hahn, 1991; Sasieni, 1971, 1989). The findings of such studies imply that for a concave or linear advertising response function, a policy of even spending is optimal, whereas for a convex function, the best practical advertising policy is one of pulsing according to which advertising is turned on-and-off in an alternating fashion. All these models, however, have ignored competitors' advertising strategies, which can have a significant impact on a firm's advertising performance (Little, 1979).

Recent studies that are relevant to the scope and purposes of this article that employ an explicitly competition-oriented focus include Mesak and Darrat (1993), Mesak and Calloway (1995, 1999), Mesak (1999), and Park and Hahn (1991). With the exception of the last study (Park & Hahn, 1991), all of the above studies employ a continuous Lanchester model and consider an infinite planning horizon. Because firms plan on spending their advertising funds over a period of time that does not usually exceed one year, the present study considers a finite planning horizon and a dynamic programming (DP) approach to determine the optimal advertising policy of a firm responding to the advertising pulsation policy of its rival.

In the marketing literature, two questions of particular significance remain only partially answered. The first is concerned with the best way of allocating advertising funds over a finite planning horizon of *n* equal consecutive time periods so that a certain performance measure is optimized in a competitive setting. The second question is whether the optimal advertising policy differs from the pulsing or even policies frequently discussed in the literature. In this paper, we attempt to apply the DP approach to address these two questions in a duopolistic framework. Additionally, the performances of the pulsing and even policies will be compared to that derived from the DP approach in a numerical setting.

MATHEMATICAL MODEL

Here we choose one of the two competitors in a duopolistic market as the focal firm, which is assumed to observe and then respond to its rival's advertising strategy. The Lanchester model delineates the causal relationship between the advertising levels of the two rivals, as well as the sales response to advertising of the focal firm. The Lanchester model has been used in several studies to examine optimal advertising policies in competitive advertising situations and is claimed to be one of the most popular models of advertising competition (Chintagunta & Vilcassim, 1994; Erickson, 1985; Horsky, 1977; Kimball, 1957; Little, 1979). Little (1979) shows that it is also one of the most realistic response models that reflect essential advertising phenomena, and it is a competitive version of the well-known Vidale–Wolfe model (Vidale & Wolfe, 1957). Park and Hahn (1991) highlight that a duopolistic Lanchester model does not constitute a severe restrictive assumption because it describes many interesting competitive situations (e.g., Coke vs. Pepsi). Here, we use a continuous version of the Lanchester model to represent the impact of competitive advertising on the sales rate of the focal firm in a duopolistic setting.

From this point on, the focal firm is referred to as firm 1 and its rival as firm 2. Let $S_i(t)$ denote the sales rate of firm i (\$/unit time) ($i = 1, 2$) at time t; let x and y stand for the advertising rates of firms 1 and 2 (\$/unit time), respectively. Then, the instantaneous change in the sales rates of the two competing firms expressed by a modified Lanchester model is as follows:

$$\mathrm{d}S_1(t)/\mathrm{d}t = -\mathrm{d}S_2(t)/\mathrm{d}t = f(x)S_2(t) - g(y)S_1(t). \tag{1}$$

The advertising response functions, $f(x)$ and $g(y)$, are assumed to take on the following two functional forms:

$$f(x) = b_1 x^{\delta_1}, \tag{2.1}$$

$$g(y) = b_2 y^{\delta_2}. \tag{2.2}$$

where b_i is the measure of advertising effectiveness of firm i ($i = 1, 2$) and δ_i the measure of the degree of convexity (concavity) of the advertising response functions $f(\cdot)$ *and* $g(\cdot)$, respectively. The assumption of a constant market of size m is plausible for mature markets and implies that

$$S_1(t) + S_2(t) = m, \quad \text{for all time } t. \tag{3}$$

The steady-state sales rates of firm i ($i = 1, 2$), $\Phi_i(x,y)$, can be obtained through equating $dS_1(t)/dt = dS_2(t)/dt = 0$, then solving Eqs. (1) and (3) simultaneously to obtain

$$\Phi_1(x,y) = m - \Phi_2(x,y) = \frac{mf(x)}{f(x)+g(y)}. \tag{4}$$

It can be shown that for rectangular or uniform advertising policies, Eq. (1) takes the following form

$$\mathrm{d}S_1(t)/\mathrm{d}t = -\mathrm{d}S_2(t)/\mathrm{d}t = [f(x)+g(y)][\Phi_1(x,y) - S_1(t)]. \tag{5}$$

In the following discussions, the planning horizon is assumed to comprise n equal consecutive time periods each of length T. Beginning from the starting point of the planning horizon, the n periods are successively denoted as period i ($i = 1, 2,\ldots,n$). Since firm 1 is not going out of business by the end of the nth period, the infinite period immediately following the planning horizon must also be considered to assess the effect of advertising efforts in previous periods (see Little (1979) for further discussion on the end effects). This infinite period is denoted as period $n+1$. Several key notations used in our mathematical model are defined below for the purpose of providing a concise exposition:

S_i is the sales rate of firm 1 at the beginning of period i;
I_i the advertising budget of firm 1 available at the beginning of period i;
I the total advertising budget of firm 1 available at the beginning of the n-period planning horizon;
x_t the advertising rate of firm 1 in period i; and
y_i the advertising rate of firm 2 in period i.

Now, let us examine firm 1's sales rate curve $q_i(t)$ in Fig. 1 for period i ($i = 1, 2,\ldots,n$) over which the advertising funds of firms 1 and 2 are assumed to be spent at the rates x_i and y_i, respectively. The solution of the differential Eq. (5) produces

$$q_i(t) = S_i\mathrm{e}^{-(f(x_i)+g(y_i))(t-(i-1)T)} + \Phi_1(x_i,y_i)(1-\mathrm{e}^{-(f(x_i)+g(y_i))(t-(i-1)T)}),$$
$$(i-1)T \le t \le iT. \tag{6}$$

Since it is assumed that no advertising of firm 1 occurs beyond the n-period planning horizon, the sales rate of firm 1 decays exponentially over period $n+1$ as time elapses. As a result of solving (5) when $x = 0$, the sales rate curve over period $n+1$ takes the following form:

$$q_\infty(t) = S_{n+1}\mathrm{e}^{-g(y)(t-nT)}, \quad t>nT. \tag{7}$$

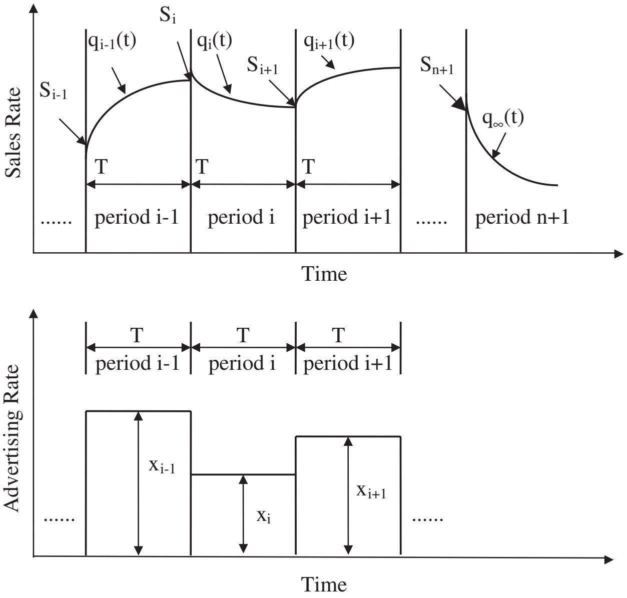


Fig. 1. Sales Response of the Focal Firm to Rectangular Pulses of Advertising.

Uniform, pulsing-maintenance, and pulsing policies are three major types of advertising pulsation policies frequently discussed in the marketing literature. They are defined below.

(1) *Uniform advertising policy (UAP)*: The firm advertises at some constant level throughout the entire planning horizon.

(2) *Advertising pulsing/maintenance policy (APMP)*: The firm alternates between high and low levels of advertising in each period throughout the planning horizon. There are two different patterns of APMP. If the firm starts with a high level of advertising, this policy type is designated as APMP-I. An APMP-II type is the one for which the firm starts with a low level of advertising.

(3) *Advertising pulsing policy* (*APP*): The firm alternates between high and zero levels of advertising in each period. APP-I and APP-II start with a high and a zero level, respectively.

It is worth mentioning at this point that for a set of alternative advertising policies that cost the same, maximizing sales revenue is equivalent to maximizing profit, as far as the ratio of cost (other than advertising expenditure) to sales revenue is constant over time and independent of these policies (refer to Mesak (1985) for a detailed discussion). Therefore, the total sales revenue accumulated over the planning horizon plus the infinite period is chosen as the performance measure of firm 1's DP and traditional advertising pulsation policies, namely UAP, APMP, and APP.

Given an advertising budget I of firm 1 at time $t = 0$ (i.e., the beginning of the planning horizon), which is assumed to be exhaustively spent over the planning horizon, the problem of optimally allocating the budget becomes one of maximizing firm 1's total sales revenue. Thus, we aim at finding the optimal advertising rate x_i^* $(i = 1, 2, \ldots ,n)$ for each period i within the planning horizon to

$$\mathrm{Max}\left\{\sum_{i=1}^{n}\int_0^T q_i(t)\,\mathrm{d}t + \int_0^\infty q_\infty(t)\,\mathrm{d}t\right\}$$

Subject to

$$\begin{gathered}\sum_{i=1}^{n} x_i T = I,\\ \text{and} \quad x_i \geq 0, \qquad i = 1, 2, \ldots, n.\end{gathered} \tag{8}$$

It is noted in the above formulation that a change in the time variable is used, so that time is set equal to zero at the beginning of each time period.

Owing to the complex nonlinear structure of the objective function in the mathematical programming model (8), the solution is extremely difficult to obtain using the ordinary nonlinear programming methods such as those based on the well-known Karush–Kuhn–Tucker (KKT) conditions and gradient search, since the equations related to the KKT conditions are difficult, if not impossible, to solve analytically for the decision variables. Thanks to the principle of decomposition, DP appears to provide an effective solution technique that meets the requirements of the optimization problem. For solution purposes, the mathematical programming model (8) is cast into a DP formulation in the next section.

DYNAMIC PROGRAMMING FORMULATION

In general, a DP formulation has six key components: (a) sequence of decision stages; (b) input state vector; (c) decision vector; (d) transition function; (e) stage return; and (f) recursive relationship. These components, shown in Fig. 2, are identified and discussed below.

The entire planning horizon is divided into a sequence of n consecutive time periods of equal length, each of which stands for a decision stage. These stages are indexed corresponding to the indices of the time periods defined in the previous section. The n-stage planning horizon plus the infinite stage (i.e., stage $n+1$) provides a framework to decompose the mathematical programming model into a sequence of smaller and computationally simpler subproblems.

The input state vector of stage i, $\boldsymbol{A_i}$ ($i = 1, 2,\ldots,n+1$), contains two elements, firm 1's initial sales rate S_i and advertising budget available I_i at the beginning of the stage. Obviously, as shown in Fig. 2, each stage's output state vector serves as the input state vector of the next stage. In particular, $\boldsymbol{A_{n+1}}$ stands for the output state vector of the last stage of the planning horizon, which is composed of the initial sales rate S_{n+1} and a zero advertising budget at the end of the planning horizon (the total advertising budget I must be exhausted over the planning horizon).

The decision vector for each stage contains only one decision variable, i.e., the rate of advertising spending in the stage, and thus, the decision vector x_i ($i = 1, 2,\ldots,n$) reduces to a scalar variable. It should also be noted that all the n decision variables are subjected to the budgetary constraint depicted in (8).

The transition function defines the linkage between the input and output state vectors of a stage and the decision made in the stage. It may be expressed as follows:

$$\boldsymbol{A}_{i+1} = t_i(\boldsymbol{A}_i)$$

where,

$$\boldsymbol{A}_{i+1} = [S_{i+1}, I_{i+1}], \quad \boldsymbol{A}_i = [S_i, I_i],$$

$$S_1 \text{ is given.}$$

$$S_{i+1} = S_i \mathrm{e}^{-(f(x_i)+g(y_i))T} + \Phi_1(x_i, y_i)(1 - \mathrm{e}^{-(f(x_i)+g(y_i))T}), \quad i = 1, 2, \ldots, n. \tag{9}$$

$$I_1 = I \text{ is given}; \quad I_{i+1} = I_i - x_i T; \quad t_i = \text{symbol of function of}, \quad i = 2, \ldots, n. \tag{10}$$

Based on the budgetary constraint (8) and the relationship (10), it can be easily shown that $I_{n+1} = 0$.

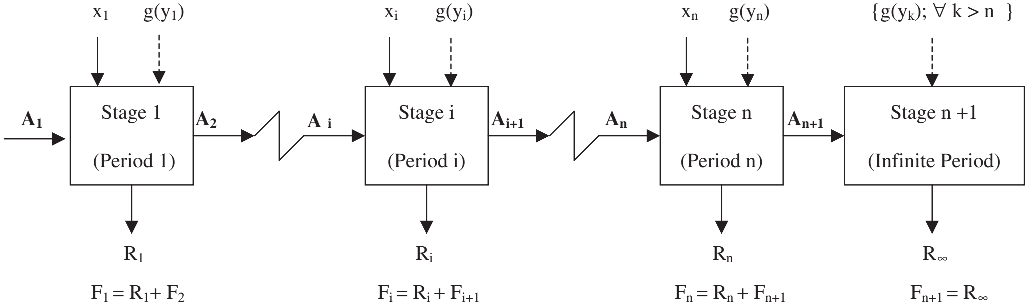


Fig. 2. Dynamic Programming Formulation of the Optimization Problem.

The sales revenue $R_i(S_i, I_i, x_i)$, the return at stage i, is a function of the input state vector $\boldsymbol{A_i} = (S_i, I_i)$ and the decision variable x_i. Specifically, for $1 \leqslant i \leqslant n$,

$$R_i(S_i, I_i, x_i) = \int_0^T q_i(t)\,\mathrm{d}t = \Phi_1(x_i, y_i)T + \frac{S_i - \Phi_1(x_i, y_i)}{f(x_i) + g(y_i)}(1 - \mathrm{e}^{-(f(x_i)+g(y_i))T}). \tag{11}$$

It is noted that $R_i(S_i, I_i, x_i)$ is conditioned by firm 2's advertising effectiveness $g(y_i)$ at stage i, a parameter beyond firm 1's control.

Next, we examine in detail firm 1's sales revenue for the infinite period following the planning horizon that ends by time period n at the end of which the sales rate of the firm is S_{n+1}. Assume that firm 2 indefinitely implements the uniform (even) policy or one of the other pulsation policies defined earlier. If firm 2 employs UAP with a constant advertising rate y_c, let $g_c = g(y_c)$; if firm 2 employs APMP alternating at two advertising rates y_h and y_l ($y_h > y_l$), let $g_h = g(y_h)$ and $g_l = g(y_l)$. It is shown in the Appendix that firm 1's sales revenue for the infinite period takes one of the following five forms:

(1) If firm 2 employs UAP indefinitely at the constant spending g_c,

$$R_\infty = \frac{S_{n+1}}{g_c}. \tag{12}$$

(2a) If firm 2 employs APMP-I indefinitely alternating between spending levels g_h and g_l,

$$R_\infty = \frac{S_{n+1}}{1 - \mathrm{e}^{-(g_h+g_l)T}}\left[\frac{1 - \mathrm{e}^{-g_h T}}{g_h} + \frac{\mathrm{e}^{-g_h T}(1 - \mathrm{e}^{-g_l T})}{g_l}\right]. \tag{13.1}$$

(2b) If firm 2 employs APMP-II indefinitely alternating between spending levels g_l and g_h,

$$R_\infty = \frac{S_{n+1}}{1 - \mathrm{e}^{-(g_h+g_l)T}}\left[\frac{1 - \mathrm{e}^{-g_l T}}{g_l} + \frac{\mathrm{e}^{-g_l T}(1 - \mathrm{e}^{-g_h T})}{g_h}\right]. \tag{13.2}$$

(3a) If firm 2 employs APP-I indefinitely alternating between spending levels g_h and 0,

$$R_\infty = S_{n+1}\left[\frac{1}{g_h} + \frac{T\mathrm{e}^{-g_h T}}{1 - \mathrm{e}^{-g_h T}}\right]. \tag{14.1}$$

(3b) If firm 2 employs APP-II indefinitely alternating between spending levels 0 and g_h,

$$R_\infty = S_{n+1}\left[\frac{1}{g_h} + \frac{T}{1 - \mathrm{e}^{-g_h T}}\right]. \tag{14.2}$$

We note in each of the above five expressions that R_∞ is conditioned by firm 2's advertising policy.

In this paper, we follow a backward induction process to formulate the recursive relationship. Starting from stage $n + 1$, the recursive relationship is described as follows:

For stage n,

$$F_n^*(S_n, I_n) = \underset{\forall x_n = I_n/T}{\mathrm{Max}}\left\{R_n(S_n, I_n, x_n) + R_\infty\right\}, \tag{15}$$

where R_∞ is stated in terms of S_n and x_n through (9) and (12), (13.1), (13.2), (14.1), or (14.2), whichever is appropriate under a given scenario.

For stage i ($i = 1, 2,\ldots,n-1$),

$$F_i^*(S_i, I_i) = \underset{\forall x_i \le I_i/T}{\mathrm{Max}}\left\{R_i(S_i, I_i, x_i) + F_{i+1}^*(S_{i+1}, I_{i+1})\right\}, \tag{16}$$

where, S_{i+1} is stated in terms of S_i and x_i through (9) and I_{i+1} in terms of I_i and x_i through (10).

The optimal solution to the DP model formulated above, x_i^* ($i = 1, 2,\ldots,n$), is functionally dependent upon the two state variables S_i and I_i and thus can be expressed as $x_i^*(S_i, I_i)$. The recursive optimization is carried out backward until the first stage is reached. At stage 1, the maximum total return (i.e., total sales revenue) $F_1^*(S_1, I_1)$ and the corresponding optimal advertising rate, $x_1^* = x_1^*(S_1, I_1)$, are determined. It is noted that $x_1^* = x_1^*(S_1, I_1)$ is a unique value due to the fact that S_1 and $I_1 = I$ are given. It is then possible to backtrack from the first stage through the succeeding stages to obtain the optimal advertising rates for all the other stages in the following manner: At stage 2, compute the optimal state pair S_2^* and I_2^* using S_1 and I_1 and x_1^* through (9) and (10) respectively and then determine the optimal advertising rate for stage 2 through the function $x_2^* = x_2^*(S_2^*, I_2^*)$. Afterwards, for stage i ($i = 3, 4,\ldots,n$), compute the optimal state pair S_i^* and I_i^* using S_{i-1}^* and I_{i-1}^* and x_{i-1}^* through (9) and (10), respectively, and then determine the optimal advertising rate for stage i through the function $x_i^* = x_i^*(S_i^*, I_i^*)$.

A NUMERICAL ILLUSTRATION

In this section, a numerical example is presented illustrating the application of DP to produce the optimal advertising schedule related to a potentially real business scenario. Let us examine a duopolistic market where the sales response of the focal firm to advertising is governed by the continuous Lanchester model (5). Consider a planning horizon that consists of four equal time periods. For illustrative purposes, assume a market potential of $m = 100$ million dollars per year. In addition, let us suppose that the firm would allocate exhaustively an advertising budget $I = 4$ million dollars over a planning horizon of one year composed of $n = 4$ equal quarters of duration $T = 0.25$ year each, while firm 2 employs an advertising policy of UAP, APMP-I, APMP-II, APP-I, or APP-II indefinitely. Consider the following advertising response function for firm 1:

$$f(x) = 0.2x^{\delta_1}. \tag{17}$$

If firm 2 employs UAP, assume that $g_c = 0.5$; if it employs APMP-I or APMP-II, let $g_h = 0.75$ and $g_l = 0.25$; if it employs APP-I or APP-II, let $g_h = 1.5$ and $g_l = 0$.

It is noted that firm 1's initial sales rate, S_1, cannot exceed the market potential m. Therefore, 11 alternative values of S_1, each being smaller than or equal to the market potential and measured in millions of dollars, are considered in the numerical example. They are given by $10k$; $k = 0, 1,\ldots,10$. In addition, eight values of δ_1, 0.25, 0.50, 0.75, 1.00, 1.25, 1.50, 1.75, 2.00 are considered to investigate the impact of the shape of firm 1's advertising response function on its optimal advertising policy.

The domain of the state variable I_i, the advertising funds available at the beginning of stage i, is discretized so that it would take on one of 21 possible values $0.05kI$; $k = 0, 1, \ldots, 20$ for $i = 1, 2, 3, 4$. The DP formulation of the maximization problem (8) and the procedure for obtaining the optimum advertising rates x_1^*, x_2^*, x_3^*, and x_4^* related to (8) follow closely the discussion in the third section.

A computing routine is developed for each of the five scenarios in which firm 2 employs UAP, APMP-I, APMP-II, APP-I or APP-II by coding in C + + both the recursive relationship characterized by (15) and (16) and the backtracking approach introduced in the previous section. One of the major features of these computing routines is that they can accommodate various alternative values of the key parameters (i.e., firm 1's initial sales rate S_1 and its shaping parameter δ_1, and firm 2's advertising effectiveness measures,

g_c, g_h, and g_l). This feature greatly facilitates the sensitivity analysis of the impact of changes in the model parameters on the behavioral patterns of firm 1's optimal advertising policy and the corresponding total returns. Furthermore, the computing routines can be easily modified for a finer discretized domain of the advertising budget at each decision stage. A personal computer that runs at 1.3 GHz with a 256 MB RAM was used to obtain the optimal solutions. All of the five developed computing routines based on the DP approach are exceptionally fast. As a whole, it took approximately 41 sec for the computing routines to produce the optimal solutions for the 440 (88×5) considered cases.

Owing to the limited space of presentation, the computational results are only partially reported in Tables 1–6. Although the following discussions are based on the partially demonstrated data, they shed interesting light on the behavioral patterns of firm 1's DP optimal advertising policy. The tables also report the total returns under firm 1's UAP, APP-I and APP-II or APMP-I and APMP-II, where the two levels of firm 1's advertising rate for the two APMP policies are assumed to be $x_h = 6$ million dollars per year and $x_l = 2$ million dollars per year. For example, given a shaping parameter $\delta_1 = 0.5$ and an initial sales rate $S_1 = 50$ million dollars per year, if firm 2 employs UAP with $g_c = 0.5$, then firm 1's DP optimal policy is found to require the advertising funds to be spent at a rate of 3.2 million dollars per year over the first quarter, 4 million dollars per year over both the second and third quarters, and 4.8 million dollars per year over the fourth quarter. As a result of implementing this advertising policy, the total sales revenue yielded is 141.666 million dollars. For the same parameters, the total returns under UAP, APMP-I and APMP-II are 141.514 million, 139.866 million and 140.721 million dollars, respectively.

As expected, it is shown in Tables 1–6 that firm 1's DP policy is superior to UAP, APMP-I, APMP-II, APP-I, and APP-II in terms of total yielded returns. In addition, the initial sales rate, S_1, the shaping parameter δ_1, and firm 2's advertising policy are found to be pivotal in dictating firm 1's DP policy. Let us first examine the cases under a concave advertising response function. All the six tables show that as the initial sales rate (S_1) increases, the advertising rate in the first quarter dictated by the DP policy decreases, while that in the fourth quarter increases in an accelerating manner.

Under a linear or convex advertising response function, all the six tables show that firm 1's DP policy requires little advertising effort to be committed in either the second or the third quarter, and at least half of the entire advertising budget to be allocated to the fourth quarter in most cases. For example, if firm 2 adopts UAP, firm 1 should not allocate any advertising funds to the third

Table 1. Total Returns of Firm 1's Advertising Policies Conditioned by Firm 2's UAP with $g_c = 0.5$.

S_1	x_1^*	x_2^*	x_3^*	x_4^*	DP*	UAP	APMP-I	APMP-II
$\delta_1 = 0.5$								
0	4	4	4	4	67.888	67.888	65.914	65.914
10	4	4	4	4	82.613	82.613	80.704	80.875
20	4	4	4	4	97.338	97.338	95.494	95.836
30	3.2	4	4	4.8	112.095	112.064	110.285	110.798
40	3.2	4	4	4.8	126.880	126.789	125.075	125.759
50	3.2	4	4	4.8	141.666	141.514	139.866	140.721
60	2.4	3.2	4.8	5.6	156.510	156.239	154.656	155.682
70	2.4	3.2	4.8	5.6	171.410	170.964	169.446	170.643
80	1.6	3.2	4.8	6.4	186.397	185.689	184.237	185.605
90	0.8	2.4	4.8	8	201.518	200.414	199.027	200.566
100	0.8	1.6	4.8	8.8	216.833	215.139	213.818	215.527
$\delta_1 = 1.0$								
0	8	0	0	8	117.852	116.637	116.439	116.439
10	7.2	0	0	8.8	128.864	127.683	127.217	127.772
20	5.6	0	0	10.4	140.150	138.730	137.995	139.105
30	4	0	0	12	151.723	149.776	148.773	150.439
40	1.6	1.6	0	12.8	163.694	160.823	159.551	161.772
50	0	1.6	0	14.4	176.108	171.869	170.329	173.105
60	0	0	0	16	188.854	182.916	181.108	184.439
70	0	0	0	16	201.680	193.963	191.886	195.772
80	0	0	0	16	214.506	205.009	202.664	207.105
90	0	0	0	16	227.332	216.056	213.442	218.439
100	0	0	0	16	240.158	227.102	224.220	229.772
$\delta_1 = 1.5$								
0	8	0	0	8	212.420	178.073	184.919	184.919
10	8	0	0	8	216.817	184.701	190.472	191.629
20	8	0	0	8	221.513	191.329	196.025	198.339
30	8	0	0	8	226.555	197.957	201.579	205.049
40	7.2	0	0	8.8	231.998	204.585	207.132	211.760
50	7.2	0	0	8.8	238.876	211.213	212.685	218.470
60	7.2	0	0	8.8	246.123	217.841	218.239	225.180
70	6.4	0	0	9.6	253.370	224.468	223.792	231.890
80	6.4	0	0	9.6	260.616	231.096	229.346	238.600
90	0	6.4	0	9.6	267.863	237.724	234.899	245.311
100	0	5.6	0	10.4	275.110	244.352	240.452	252.021

Table 2. Total Returns of Firm 1's Advertising Policies Conditioned by Firm 2's UAP with $g_c = 0.5$.

S_1	x_1^*	x_2^*	x_3^*	x_4^*	DP*	UAP	APP-I	APP-II
$\delta_1 = 0.5$								
0	4	4	4	4	67.888	67.888	50.180	50.180
10	4	4	4	4	82.613	82.613	66.019	66.508
20	4	4	4	4	97.338	97.338	81.857	82.835
30	3.2	4	4	4.8	112.095	112.064	97.695	99.162
40	3.2	4	4	4.8	126.880	126.789	113.534	115.490
50	3.2	4	4	4.8	141.666	141.514	129.372	131.817
60	2.4	3.2	4.8	5.6	156.510	156.239	145.211	148.145
70	2.4	3.2	4.8	5.6	171.410	170.964	161.049	164.472
80	1.6	3.2	4.8	6.4	186.397	185.689	176.887	180.799
90	0.8	2.4	4.8	8	201.518	200.414	192.725	197.126
100	0.8	1.6	4.8	8.8	216.833	215.139	208.564	213.454
$\delta_1 = 1.0$								
0	8	0	0	8	117.852	116.637	115.843	115.843
10	7.2	0	0	8.8	128.864	127.683	126.370	127.483
20	5.6	0	0	10.4	140.150	138.730	136.897	139.123
30	4	0	0	12	151.723	149.776	147.424	150.763
40	1.6	1.6	0	12.8	163.694	160.823	157.951	162.403
50	0	1.6	0	14.4	176.108	171.869	168.478	174.043
60	0	0	0	16	188.854	182.916	179.005	185.683
70	0	0	0	16	201.680	193.963	189.532	197.323
80	0	0	0	16	214.506	205.009	200.059	208.963
90	0	0	0	16	227.332	216.056	210.585	220.603
100	0	0	0	16	240.158	227.102	221.112	232.244
$\delta_1 = 1.5$								
0	8	0	0	8	212.420	178.073	203.815	203.815
10	8	0	0	8	216.817	184.701	207.695	209.589
20	8	0	0	8	221.513	191.329	211.576	215.364
30	8	0	0	8	226.555	197.957	215.456	221.138
40	7.2	0	0	8.8	231.998	204.585	219.337	226.913
50	7.2	0	0	8.8	238.876	211.213	223.217	232.688
60	7.2	0	0	8.8	246.123	217.841	227.098	238.462
70	6.4	0	0	9.6	253.370	224.468	230.978	244.237
80	6.4	0	0	9.6	260.616	231.096	234.858	250.011
90	0	6.4	0	9.6	267.863	237.724	238.739	255.786
100	0	5.6	0	10.4	275.110	244.352	242.619	261.560

Table 3. Total Returns of Firm 1's Advertising Policies Conditioned by Firm 2's APMP-I with $g_h = 0.75$ and $g_l = 0.25$.

S_1	x_1^*	x_2^*	x_3^*	x_4^*	DP*	UAP	APMP-I	APMP-II
$\delta_1 = 0.5$								
0	4	4	4	4	67.905	67.905	65.935	65.926
10	4	4	4	4	82.181	82.181	80.275	80.430
20	4	4	4	4	96.456	96.456	94.614	94.934
30	3.2	4	4	4.8	110.756	110.732	108.954	109.438
40	3.2	4	4	4.8	125.090	125.008	123.294	123.942
50	3.2	4	4	4.8	139.424	139.283	137.634	138.445
60	2.4	3.2	4.8	5.6	153.795	153.559	151.973	152.949
70	2.4	3.2	4.8	5.6	168.241	167.834	166.313	167.452
80	1.6	3.2	4.8	6.4	182.739	182.110	180.653	181.956
90	1.6	2.4	4.8	7.2	197.336	196.385	194.992	196.460
100	0.8	2.4	4.8	8	212.151	210.661	209.332	210.963
$\delta_1 = 1.0$								
0	8	0	0	8	117.839	116.654	116.529	116.392
10	7.2	0	0	8.8	128.479	127.364	126.982	127.377
20	5.6	0	0	10.4	139.335	138.074	137.435	138.363
30	4	0.8	0	11.2	150.497	148.784	147.887	149.348
40	1.6	1.6	0	12.8	161.995	159.494	158.340	160.333
50	0	2.4	0	13.6	173.918	170.204	168.792	171.318
60	0	0	0	16	186.156	180.914	179.245	182.304
70	0	0	0	16	198.585	191.625	189.698	193.289
80	0	0	0	16	211.015	202.335	200.150	204.274
90	0	0	0	16	223.445	213.045	210.603	215.259
100	0	0	0	16	235.874	223.755	221.055	226.245
$\delta_1 = 1.5$								
0	8.8	0	0	7.2	212.100	178.043	185.327	184.515
10	8	0	0	8	216.173	184.471	190.719	191.016
20	7.2	0	0	8.8	220.509	190.899	196.110	197.518
30	6.4	0	0	9.6	225.141	197.326	201.501	204.019
40	5.6	0	0	10.4	230.112	203.754	206.893	210.521
50	4.8	0	0	11.2	235.474	210.182	212.284	217.022
60	0	0	0	16	242.298	216.610	217.675	223.524
70	0	0	0	16	249.317	223.038	223.067	230.025
80	0	0	0	16	256.337	229.466	228.458	236.527
90	0	0	0	16	263.356	235.894	233.849	243.028
100	0	0	0	16	270.375	242.322	239.241	249.530

Table 4. Total Returns of Firm 1's Advertising Policies Conditioned by Firm 2's APMP-II with $g_h = 0.75$ and $g_l = 0.25$.

S_1	x_1^*	x_2^*	x_3^*	x_4^*	DP*	UAP	APMP-I	APMP-II
$\delta_1 = 0.5$								
0	4	4	4	4	67.915	67.915	65.935	65.943
10	4	4	4	4	83.109	83.109	81.195	81.382
20	4	4	4	4	98.302	98.302	96.455	96.820
30	3.2	4	4	4.8	113.535	113.497	111.716	112.259
40	3.2	4	4	4.8	128.791	128.690	126.976	127.698
50	3.2	3.2	4	5.6	144.055	143.884	142.236	143.136
60	2.4	3.2	4.8	5.6	159.384	159.078	157.497	158.575
70	2.4	3.2	4	6.4	174.764	174.272	172.757	174.013
80	1.6	3.2	4.8	6.4	190.254	189.465	188.017	189.451
90	0.8	2.4	4.8	8	205.933	204.659	203.277	204.890
100	0	1.6	4.8	9.6	221.913	219.853	218.537	220.328
$\delta_1 = 1.0$								
0	8	0	0	8	117.941	116.695	116.423	116.561
10	6.4	0	0	9.6	129.373	128.092	127.541	128.257
20	5.6	0	0	10.4	141.070	139.490	138.658	139.953
30	3.2	0.8	0	12	153.115	150.887	149.776	151.649
40	1.6	0.8	0	13.6	165.563	162.284	160.893	163.346
50	0	0.8	0	15.2	178.500	173.682	172.011	175.042
60	0	0	0	16	191.727	185.079	183.128	186.738
70	0	0	0	16	204.966	196.476	194.246	198.435
80	0	0	0	16	218.205	207.873	205.363	210.131
90	0	0	0	16	231.444	219.271	216.481	221.827
100	0	0	0	16	244.683	230.668	227.599	233.524
$\delta_1 = 1.5$								
0	7.2	0	0	8.8	212.992	178.219	184.629	185.442
10	7.2	0	0	8.8	217.741	185.055	190.352	192.370
20	6.4	0	0	9.6	222.850	191.891	196.074	199.298
30	4.8	0	0	11.2	228.337	198.727	201.796	206.226
40	0	0	0	16	235.181	205.564	207.518	213.154
50	0	0	0	16	242.665	212.400	213.240	220.082
60	0	0	0	16	250.149	219.236	218.962	227.010
70	0	0	0	16	257.633	226.073	224.684	233.938
80	0	0	0	16	265.118	232.909	230.406	240.866
90	0	0	0	16	272.602	239.745	236.129	247.794
100	0	0	0	16	280.086	246.582	241.851	254.721

Table 5. Total Returns of Firm 1's Advertising Policies Conditioned by Firm 2's APP-I with $g_h = 1.5$ and $g_l = 0$.

S_1	x_1^*	x_2^*	x_3^*	x_4^*	DP*	UAP	APP-I	APP-II
$\delta_1 = 0.5$								
0	4	4	4	4	45.927	45.927	33.772	33.843
1	4	4	4	4	55.200	55.200	43.601	44.060
20	4	4	4	4	64.473	64.473	53.430	54.276
30	3.2	4	4	4.8	73.779	73.745	63.258	64.492
40	3.2	4	4	4.8	83.098	83.018	73.087	74.708
50	3.2	3.2	4	5.6	92.420	92.291	82.915	84.924
60	2.4	3.2	4.8	5.6	101.786	101.564	92.744	95.140
70	2.4	3.2	4.8	5.6	111.186	110.836	102.573	105.357
80	1.6	3.2	4.8	6.4	120.625	120.109	112.402	115.573
90	1.6	2.4	4.8	7.2	130.110	129.382	122.230	125.789
100	0.8	2.4	4.8	8	139.714	138.654	132.059	136.006
$\delta 1 = 1.0$								
0	8	0.8	0	7.2	80.709	79.776	79.259	78.896
10	7.2	0.8	0	8	87.810	86.997	86.064	86.591
20	5.6	0.8	0	9.6	95.096	94.217	92.869	94.286
30	3.2	2.4	0	10.4	102.609	101.439	99.673	101.981
40	1.6	3.2	0	11.2	110.390	108.660	106.479	109.676
50	0	3.2	0	12.8	118.485	115.880	113.283	117.371
60	0	1.6	0.8	13.6	126.759	123.101	120.088	125.066
70	0	0.8	0.8	14.4	135.195	130.322	126.893	132.761
80	0	0.8	0	15.2	143.716	137.543	133.697	140.456
90	0	0.8	0	15.2	152.238	144.764	140.502	148.151
100	0	0.8	0	15.2	160.759	151.985	147.307	155.846
$\delta_1 = 1.5$								
0	8.8	0.8	0	6.4	146.369	124.164	145.016	141.801
10	8	0.8	0	7.2	149.518	128.843	147.869	146.202
20	7.2	0.8	0	8	152.840	133.522	150.722	150.604
30	6.4	0.8	0	8.8	156.369	138.202	153.575	155.005
40	5.6	0.8	0	9.6	160.153	142.881	156.428	159.407
50	0	5.6	0	10.4	164.826	147.560	159.281	163.809
60	0	4.8	0	11.2	169.783	152.239	162.134	168.210
70	0	4	0	12	174.908	156.918	164.987	172.612
80	0	0.8	0	15.2	180.297	161.597	167.840	177.014
90	0	0.8	0	15.2	186.103	166.277	170.693	181.415
100	0	0.8	0	15.2	191.909	170.956	173.546	185.817

Table 6. Total Returns of Firm 1's Advertising Policies Conditioned by Firm 2's APP-II with $g_h = 1.5$ and $g_l = 0$.

S_1	x_1^*	x_2^*	x_3^*	x_4^*	DP*	UAP	APP-I	APP-II
$\delta_1 = 0.5$								
0	4	4	4	4	45.951	45.951	33.843	33.772
10	4	4	4	4	57.125	57.125	45.675	46.101
20	3.2	4	4	4.8	68.306	68.299	57.506	58.430
30	3.2	4	4	4.8	79.537	79.473	69.337	70.758
40	3.2	4	4	4.8	90.768	90.647	81.168	83.087
50	2.4	3.2	4.8	5.6	102.052	101.821	92.999	95.415
60	2.4	3.2	4.8	5.6	113.381	112.995	104.831	107.744
70	1.6	3.2	4.8	6.4	124.791	124.169	116.662	120.073
80	0.8	2.4	4.8	8	136.291	135.343	128.493	132.402
90	0.8	2.4	4.8	8	147.932	146.518	140.324	144.730
100	0.8	1.6	4.8	8.8	159.641	157.692	152.156	157.059
$\delta_1 = 1.0$								
0	7.2	0	0.8	8	80.915	79.872	78.896	79.259
10	6.4	0	0.8	8.8	89.654	88.567	87.059	88.564
20	4.8	0	0.8	10.4	98.692	97.262	95.221	97.869
30	2.4	0.8	0.8	12	108.066	105.958	103.385	107.174
40	0.8	0.8	0.8	13.6	117.869	114.654	111.548	116.479
50	0.8	0	0.8	14.4	127.986	123.349	119.711	125.783
60	0.8	0	0.8	14.4	138.134	132.044	127.873	135.088
70	0.8	0	0.8	14.4	148.281	140.740	136.036	144.393
80	0.8	0	0.8	14.4	158.429	149.435	144.199	153.697
90	0.8	0	0.8	14.4	168.577	158.130	152.362	163.002
100	0.8	0	0.8	14.4	178.724	166.826	160.525	172.307
$\delta_1 = 1.5$								
0	6.4	0	0.8	8.8	147.820	124.548	141.801	145.016
10	6.4	0	0.8	8.8	152.377	130.170	145.172	150.369
20	5.6	0	0.8	9.6	157.288	135.791	148.543	155.722
30	4	0	0.8	11.2	162.706	141.412	151.914	161.075
40	0.8	0	0.8	14.4	169.448	147.033	155.285	166.428
50	0.8	0	0.8	14.4	176.455	152.655	158.656	171.781
60	0.8	0	0.8	14.4	183.462	158.276	162.027	177.134
70	0.8	0	0.8	14.4	190.469	163.897	165.398	182.487
80	0.8	0	0.8	14.4	197.476	169.519	168.769	187.840
90	0.8	0	0.8	14.4	204.484	175.140	172.140	193.193
100	0.8	0	0.8	14.4	211.491	180.761	175.511	198.546

quarter. In particular, given the linear advertising response function, firm 1 should concentrate all its advertising efforts over the fourth quarter for an S_1 of at least 60 million dollars per year. If firm 2 employs APMP-I or APMP-II, firm 1 should not spend even a penny of the advertising budget over the second and third quarters given the convex advertising response function. If firm 2 employs APP-I and APP-II, firm 1 should not commit any advertising efforts in the third quarter and the second quarter, respectively.

It is noted in Tables 3–6 that, in most cases, firm 1's DP policy yields a higher total return when firm 2 employs APMP-II or APP-II; however, if firm 2 switches from APMP-II to APMP-I or from APP-II to APP-I, the return yielded by firm 1's DP policy will be lower. Accordingly, firm 2, in general, can reduce the effectiveness of firm 1's DP policy for the considered scenarios by switching from APMP-II to APMP-I or from APP-II to APP-I, especially when the initial sales rate of firm 1 is high.

Also, some interesting comparative findings related to the traditional advertising pulsation policies of UAP, APMP-I, APMP-II, APP-I and APP-II are observed from Tables 1–6 and reported below.

(1) For firm 1 responding to its rival's UAP, APMP-I, APMP-I, APP-I or APP-II, if its advertising response function is convex, firm 1's APMP-II and APP-II dominate its UAP (Tables 1–6).
(2) For firm 1 responding to its rival's UAP and for all shapes of the advertising response function, firm 1's APMP-I and APP-I are not as effective as its APMP-II and APP-II counterparts, respectively, with the exception of the case for $S_1 = 0$ (Tables 1 and 2).
(3) For firm 1 responding to its rival's APMP-I, neither firm 1's APMP-I nor its APMP-II consistently dominates the other (Table 3).
(4) For firm 1 responding to its rival's APMP-II and for all shapes of the advertising response function, firm 1's APMP-II dominates its APMP-I (Table 4).
(5) For firm 1 responding to its rival's APP-I or APP-II, neither firm 1's APP-I nor its APP-II consistently dominates the other (Tables 5 and 6).

SUMMARY AND CONCLUSIONS

In this paper, a continuous version of the Lanchester model is employed to depict the relationship between the focal firm's sales response to competing advertising efforts in a duopolistic market. Using the DP approach, we address the problem of optimally allocating an advertising budget by the focal firm over a finite planning horizon to maximize total sales revenue related to the advertising efforts. A mathematical programming model and the related DP formulation are analytically developed to find the optimum advertising policy for the focal firm facing competitive advertising. In addition to the above main methodological contribution, a numerical example is presented to illustrate the application of the proposed DP approach. Based on the specified form of the advertising response function and the employed discretized scheme, the major findings of the numerical investigation are summarized below:

(1) Under a concave or linear advertising response function, the focal firm's dynamic programming (DP) policy dominates its Uniform Advertising Policy (UAP) irrespective of the advertising policy employed by the competition, as expected.
(2) Under a convex advertising response function, the focal firm's dynamic programming (DP) policy dominates its Advertising/Maintenance Pulsing Policy (APMP) and Advertising Pulsing Policy (APP) irrespective of the advertising policy employed by the competition, as expected.

The DP algorithm presented in this paper is developed for a general planning horizon comprising *n* equal consecutive time periods. It has been demonstrated that problems with a realistic size can be efficiently solved on a microcomputer. It is expected that microcomputer-based approaches, such as the one presented in this paper, could be used to provide the focal firm with a competitive edge in a duopolistic environment.

It is important to point out the limitations of this paper and suggest several possibilities for future research. First, in this exploratory study we focus on a firm responding to its rival's advertising pulsation strategy. For a comprehensive and perhaps a much more challenging treatment, the situation may be considered from a game theory perspective. Second, advertising spending is the only decision variable in the modeling framework. Incorporating other marketing mix variables such as price would be a possible extension. Third, sales response of the focal firm to advertising is modeled for a duopolistic market. Modeling such a response in an oligopolistic setting could be a plausible research topic in the future.

REFERENCES

Chintagunta, P. K., & Vilcassim, N. J. (1994). Marketing investment decisions in a dynamic duopoly: A model and empirical analysis. *International Journal of Research in Marketing*, *11*, 287–306.

Erickson, G. M. (1985). A model of advertising competition. *Journal of Marketing Research*, *22*(August), 297–304.

Feinberg, F. M. (1988). *Pulsing policies for aggregate advertising models*. Unpublished PhD Dissertation. Massachusetts Institute of Technology.

Hahn, M., & Hyun, J. S. (1991). Advertising cost interactions and the optimality of pulsing. *Management Science*, *37*, 157–169.

Horsky, D. (1977). An empirical analysis of the optimal advertising policy. *Management Science*, *23*, 1037–1049.

Kimball, G. E. (1957). Some industrial applications of military operations research methods. *Operations Research*, *5*(April), 201–204.

Little, J. D. C. (1979). Aggregate advertising models: The state of the art. *Operations Research*, *27*(July–August), 627–667.

Luhmer, A., Steindl, A., Feichtinger, G., Hartl, R. F., & Sorger, G. (1988). ADPULS in continuous time. *European Journal of Operations Research*, *34*, 171–177.

Mahajan, V., & Muller, E. (1986). Advertising pulsing policies for generating awareness for new products. *Marketing Science*, *5*(2), 89–106.

Mesak, H. I. (1985). On modeling advertising pulsing decisions. *Decision Sciences*, *16*(1), 25–42.

Mesak, H. I. (1999). On the generalizability of advertising pulsation monopoly results to an oligopoly. *European Journal of Operational Research*, *117*, 429–449.

Mesak, H. I., & Calloway, J. A. (1995). A pulsing model of advertising competition: A game theoretic approach, part a-theoretical foundation. *European Journal of Operational Research*, *86*, 231–248.

Mesak, H. I., & Calloway, J. A. (1999). Hybrid subgames and copycat games in a pulsing model of advertising competition. *Journal of the Operational Research Society*, *50*, 837–849.

Mesak, H. I., & Darrat, A. F. (1992). On comparing alternative advertising policies of pulsation. *Decision Sciences*, *23*(3), 541–559.

Mesak, H. I., & Darrat, A. F. (1993). A competitive advertising model: Some theoretical and empirical results. *Journal of the Operational Research Society*, *44*(5), 491–502.

Park, S., & Hanh, M. (1991). Pulsing in a discrete model of advertising competition. *Journal of Marketing Research*, *28*(November), 397–405.

Sasieni, M. W. (1971). Optimal advertising expenditures. *Management Science*, *18*, 64–72.

Sasieni, M. W. (1989). Optimal advertising strategies. *Marketing Science*, *8*(4), 358–370.

Vidale, M. L., & Wolfe, H. B. (1957). An operations research study of sales response to advertising. *Operations Research*, *5*, 370–381.

APPENDIX

Derivation of Firm 1's Sales Revenue for the Infinite Period Following the Planning Horizon

Note that in the infinite period, firm 1 shuts down advertising, and thus $f(x) = f(0) = 0$.

Case 1. Assume that firm 2 employs UAP throughout a planning horizon of n equal periods and the ensuing infinite period. Suppose $g_c = g(y_c)$, where

y_c stands for the constant advertising rate of firm 2. Recall that firm 1's sales rate at the beginning of the infinite period is denoted as S_{n+1}. Firm 1's sales rate curve over the infinite period takes the form of (7) and as a result, its sale revenue for the infinite period is given by

$$R_\infty = \lim_{\tau\to+\infty}\int_0^\tau q_\infty(t)\,\mathrm{d}t = \lim_{\tau\to+\infty}\frac{S_{n+1}}{g_c}(1-\mathrm{e}^{-g_h\tau}) = \frac{S_{n+1}}{g_c}. \tag{A1}$$

(In the above formulation a change in the time variable is used, so that time is set equal to zero at the beginning of the infinite period.)

Case 2.1. Assume that firm 2 employs APMP-I throughout a planning horizon of n equal periods of duration T and the ensuing infinite period. Suppose $g_h = g(y_h)$ and $g_l = g(y_l)$, where y_h and y_l are the high and low levels of the advertising rate between which firm 2 alternates following APMP-I. For any integer $k \geqslant 1$, firm 1's sales rate curve follows Eq. (6) over periods n+2k-1 and n+2k of duration T. As illustrated in Fig. 1, the sales rates at the beginnings of periods n+2k-1 and n+2k, S_{n+2k-1} and S_{n+2k}, can be expressed in terms of S_{n+1} using (6) in a stepwise fashion. As a result, firm 1's sales revenue for periods n+2k–1 and n+2k are given by

$$\begin{aligned} R_{n+2k-1} &= \int_0^T q_{n+2k-1}(t)\,\mathrm{d}t = \frac{S_{n+2k-1}}{g_h}(1-\mathrm{e}^{-g_hT}) \\ &= \frac{S_{n+1}}{g_h}(1-^{-g_hT})\mathrm{e}^{-(k-1)(g_h+g_l)T}, \end{aligned}$$

$$\begin{aligned} R_{n+2k} &= \int_0^T q_{n+2k}(t)\,\mathrm{d}t = \frac{S_{n+2k}}{g_l}(1-\mathrm{e}^{-g_lT}) \\ &= \frac{S_{n+1}}{g_h}(1-\mathrm{e}^{-g_hT})\mathrm{e}^{-g_hT}\mathrm{e}^{-(k-1)(g_h+g_l)T}. \end{aligned}$$

(In the above formulations a change in the time variable is used, so that time is set equal to zero at the beginning of each time period. This change in the time variable is also used in *Cases 3.1 and 3.2*.) Accordingly, firm 1's sale revenue for the infinite period is given by

$$\begin{aligned} R_\infty &= \sum_{k=1}^{\infty} R_{n+2k-1} + \sum_{k=1}^{\infty} R_{n+2k} \\ &= \frac{S_{n+1}}{1-\mathrm{e}^{-(g_h+g_l)T}}\left[\frac{1-^{-g_hT}}{g_h} + \frac{\mathrm{e}^{-g_hT}(1-\mathrm{e}^{-g_lT})}{g_l}\right]. \end{aligned} \tag{A2}$$

Case 2.2. Assume that firm 2 employs APMP-II throughout a planning horizon of n equal periods of duration T and the ensuing infinite period. Suppose $g_h = g(y_h)$ and $g_l = g(y_l)$ where y_h and y_l are the high and low levels of the advertising rate between which firm 2 alternates following APMP-II. In terms of g_h and g_l, *Case 2.2* is the mirrored image of *Case 2.1*, and therefore, firm 1's sale revenue for the infinite period is given by

$$R_\infty = \frac{S_{n+1}}{1 - \mathrm{e}^{-(g_h+g_l)T}}\left[\frac{1 - \mathrm{e}^{-g_lT}}{g_l} + \frac{\mathrm{e}^{-g_lT}(1 - \mathrm{e}^{-g_hT})}{g_h}\right]. \tag{A3}$$

Case 3.1. Assume that firm 2 employs APP-I throughout a planning horizon of n equal periods of duration T and the ensuing infinite period. Suppose $g_h = g(y_h)$ and $g_l = g(y_l) = 0$ where y_h and $y_l = 0$ are the high and zero levels of the advertising rate between which firm 2 alternates following APP-I. For any integer $k \geqslant 1$, firm 1's sales rate curve follows Eq. (6) over periods $n+2k$-*1* and $n+2k$ of duration T. Using the approach employed in *Case 2.1*, the sales rates at the beginnings of periods $n+2k$-*1* and $n+2k$, S_{n+2k-1} and S_{n+2k}, can be expressed in terms of $S_{n+1.}$ As a result, firm 1's sales revenue for periods $n+2k$–*1* and $n+2k$ are given by

$$\begin{aligned} R_{n+2k-1} &= \int_0^T q_{n+2k-1}(t)\,\mathrm{d}t = \frac{S_{n+2k-1}}{g_h}(1 - \mathrm{e}^{-g_hT}) \\ &= \frac{S_{n+1}}{g_h}(1 - \mathrm{e}^{-g_hT})\mathrm{e}^{-(k-1)g_hT}, \end{aligned}$$

$$R_{n+2k} = \int_0^T q_{n+2k}(t)\,\mathrm{d}t = S_{n+2k}T = S_{n+1}T\mathrm{e}^{-kg_hT}.$$

Accordingly, firm 1's sale revenue for the infinite period is given by

$$R_\infty = \sum_{k=1}^{\infty} R_{n+2k-1} + \sum_{k=1}^{\infty} R_{n+2k} = S_{n+1}\left[\frac{1}{g_h} + \frac{T\mathrm{e}^{-g_hT}}{1 - \mathrm{e}^{-g_hT}}\right]. \tag{A4}$$

Case 3.2. Assume that firm 2 employs APP-II throughout a planning horizon of n equal periods of duration T and the ensuing infinite period. Suppose $g_h = g(y_h)$ and $g_l = g(y_l) = 0$ where y_h and $y_l = 0$ are the high and zero levels of the advertising rate between which firm 2 alternates following APP-II. For any integer $k \geqslant 1$, firm 1's sales rate curve follows Eq. (6) over periods n+2k-*1* and n+2k of duration T. Using the approach employed in

Case 2.1, the sales rates at the beginnings of periods n+2k-*1* and n+2k, S_{n+2k-1} and S_{n+2k}, can be expressed in terms of S_{n+1}. As a result, firm 1's sales revenue for periods n+2k–*1* and n+2k are given by

$$R_{n+2k-1} = \int_0^T q_{n+2k-1}(t)\,\mathrm{d}t = S_{n+2k-1}T = S_{n+1}T\mathrm{e}^{-(k-1)g_hT},$$

$$R_{n+2k} = \int_0^T q_{n+2k}(t)\,\mathrm{d}t = \frac{S_{n+2k}}{g_h}(1-\mathrm{e}^{-g_hT}) = \frac{S_{n+1}}{g_h}(1-\mathrm{e}^{-g_hT})e^{-(k-1)g_hT}.$$

Accordingly, firm 1's sale revenue for the infinite period is given by

$$R_\infty = \sum_{k=1}^{\infty} R_{n+2k-1} + \sum_{k=1}^{\infty} R_{n+2k} = S_{n+1}\left[\frac{1}{g_h} + \frac{T}{1-\mathrm{e}^{-g_hT}}\right]. \qquad \text{(A5)}$$