Signals and Communication Technology

Gabriele Moser
Josiane Zerubia *Editors*

Mathematical Models for Remote Sensing Image Processing

Models and Methods for the Analysis of 2D Satellite and Aerial Images

Springer

Signals and Communication Technology

More information about this series at http://www.springer.com/series/4748

Gabriele Moser · Josiane Zerubia
Editors

Mathematical Models for Remote Sensing Image Processing

Models and Methods for the Analysis of 2D Satellite and Aerial Images

Springer

Editors
Gabriele Moser
University of Genoa
Genoa
Italy

Josiane Zerubia
Université Côte d'Azur
Inria
Sophia Antipolis Cedex
France

ISSN 1860-4862 ISSN 1860-4870 (electronic)
Signals and Communication Technology
ISBN 978-3-319-66328-9 ISBN 978-3-319-66330-2 (eBook)
https://doi.org/10.1007/978-3-319-66330-2

Library of Congress Control Number: 2017951438

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

To the memory of my mother, Jeanne, and my mentor, Telma, who both passed away early 2014.

Josiane Zerubia

To the memory of my grandmother, who is so sorely missed.
To all the readers. Especially the math nerds.

Gabriele Moser

Preface

This book is framed in the context of remote sensing for Earth observation and focuses on the research field of mathematical models and methodologies for the analysis of two-dimensional remote sensing images. The objective is twofold. First, the book is intended to conduct a broad analysis of the field of applied mathematics for two-dimensional remote sensing image interpretation, encompassing passive and active sensors, hyperspectral images, synthetic aperture radar (SAR), interferometric SAR, and polarimetric SAR data. Second, this book is meant to also discuss very topical and advanced subjects, which involve various types of remote sensing data (e.g., very high-resolution imagery, multiangular or multiresolution data, and satellite image time series) or of processing and learning methodologies (e.g., probabilistic graphical models, hierarchical image representations, kernel machines, data fusion, and compressive sensing) that are currently of primary importance in the Earth observation area.

The book is organized into ten chapters. The first one is introductory. It is aimed at recalling basic notions and terminology, and at providing an overview of the most prominent families of mathematical models and techniques for remote sensing image interpretation. All the other chapters are devoted either to specific typologies of remote sensing images, along with their data analysis challenges, or to individual methodological areas. Each of these nine chapters presents both a detailed review of the previous literature on the related subject and a methodological and experimental discussion of, at least, two advanced mathematical approaches to information extraction from remote sensing images. This organization is kept consistently throughout the book and allows both tutorial information and multiple on-the-edge developments to be covered on each topic.

The chapters are written and organized so that they contribute to the book as a whole and are self-consistent. Each chapter is authored by research scientists from, at least, two distinct institutions to take benefit from multiple professional experiences and perspectives on each subject. The chapter co-authors are highly prominent research scientists in the remote sensing and image processing fields. In these fields, many of them have been serving as Editors in Chief or Associate Editors in the editorial boards of prestigious international journals and are actively involved in

international scientific societies (e.g., the IEEE Geoscience and Remote Sensing Society and the IEEE Signal Processing Society). The book is intended to be used as a reference, to graduate and doctoral students and to remote sensing scientists and practitioners, as well as possibly as a textbook.

We would like to take the opportunity to acknowledge and sincerely thank all the chapter co-authors for their excellent contributions to this collective effort, their insight, and their invaluable feedback. We also wish to thank the Springer staff for their help and support along the editing and publishing process.

Genoa, Italy — Gabriele Moser
Sophia Antipolis, France — Josiane Zerubia
May 2017

Contents

1 **Mathematical Models and Methods for Remote Sensing Image Analysis: An Introduction** 1
Gabriele Moser, Josiane Zerubia, Sebastiano B. Serpico and Jon A. Benediktsson

2 **Models for Hyperspectral Image Analysis: From Unmixing to Object-Based Classification** 37
Emmanuel Maggiori, Antonio Plaza and Yuliya Tarabalka

3 **Very High Spatial Resolution Optical Imagery: Tree-Based Methods and Multi-temporal Models for Mining and Analysis** 81
Fabio Pacifici, Georgios K. Ouzounis, Lionel Gueguen, Giovanni Marchisio and William J. Emery

4 **Very-High-Resolution and Interferometric SAR: Markovian and Patch-Based Non-local Mathematical Models** 137
Charles-Alban Deledalle, Loïc Denis, Giampaolo Ferraioli, Vito Pascazio, Gilda Schirinzi and Florence Tupin

5 **Polarimetric SAR Modelling: Mellin Kind Statistics and Time-Frequency Analysis** 191
Torbjørn Eltoft, Laurent Ferro-Famil, Stian N. Anfinsen and Anthony P. Doulgeris

6 **Remote Sensing Data Fusion: Guided Filter-Based Hyperspectral Pansharpening and Graph-Based Feature-Level Fusion** 243
Wenzhi Liao, Jocelyn Chanussot and Wilfried Philips

7 Remote Sensing Data Fusion: Markov Models and Mathematical Morphology for Multisensor, Multiresolution, and Multiscale Image Classification 277
Jon A. Benediktsson, Gabriele Cavallaro, Nicola Falco, Ihsen Hedhli, Vladimir A. Krylov, Gabriele Moser, Sebastiano B. Serpico and Josiane Zerubia

8 Change Detection in Multitemporal Images Through Single- and Multi-scale Approaches 325
Bruno Aiazzi, Francesca Bovolo, Lorenzo Bruzzone, Andrea Garzelli, Davide Pirrone and Claudia Zoppetti

9 Satellite Image Time Series: Mathematical Models for Data Mining and Missing Data Restoration 357
Nicolas Méger, Edoardo Pasolli, Christophe Rigotti, Emmanuel Trouvé and Farid Melgani

10 Advances in Kernel Machines for Image Classification and Biophysical Parameter Retrieval 399
Devis Tuia, Michele Volpi, Jochem Verrelst and Gustau Camps-Valls

Chapter 1
Mathematical Models and Methods for Remote Sensing Image Analysis: An Introduction

Gabriele Moser, Josiane Zerubia, Sebastiano B. Serpico and Jon A. Benediktsson

Abstract The current progress of remote sensing systems, based on airborne and spaceborne platforms and involving active and passive sensors, provides an unprecedented wealth of information about the Earth surface for environmental monitoring, sustainable resource management, disaster prevention, emergency response, and defense. In this framework, mathematical models for image processing and analysis play fundamental roles. Effectively exploiting the potential conveyed by the availability of remote sensing data requires automatic or semi-automatic techniques capable of suitably characterizing and extracting thematic information of interest while minimizing the need for user intervention. The current development of mathematical models and methods for image processing and computer vision allows multiple remote sensing information extraction problems to be addressed successfully, accurately, and efficiently. In this introductory chapter, first, general characteristics of sensors and systems for Earth observation are summarized to define the basic terminology that will be used consistently throughout the book. Remote sensing image acquisition through passive and active sensors on-board spaceborne and airborne platforms is recalled together with the basic concepts of spatial, spectral, temporal, and radiometric resolution. Then, an overview of the main families of mathematical models and methods within the scientific field of two-dimensional remote sensing image processing is presented. The overall structure and organization of the book are also described.

G. Moser · S.B. Serpico (✉)
University of Genoa, Via Opera Pia 11a, 16145 Genoa, Italy
e-mail: sebastiano.serpico@unige.it

G. Moser
e-mail: gabriele.moser@unige.it

J. Zerubia
Université Côte d'Azur, Inria, BP 93, 2004, route des Lucioles, 06902 Sophia Antipolis Cedex, France
e-mail: josiane.zerubia@inria.fr

J.A. Benediktsson
University of Iceland, Saemundargotu 2 - 101, Reykjavik, Iceland
e-mail: benedikt@hi.is

G. Moser and J. Zerubia (eds.), *Mathematical Models for Remote Sensing Image Processing*, Signals and Communication Technology,
https://doi.org/10.1007/978-3-319-66330-2_1

1.1 Introduction

The *remote sensing* discipline studies the instrumentation, the models, the methods, the computational algorithms, and the software architectures that address the objective of retrieving information about a given "object" or "entity" by collecting and processing observations without physical contact with the object or entity itself [169]. In these terms, the definition is extremely general and encompasses very diverse application scenarios including the extraction of environmental information from data collected by satellite or aerial sensors [98], the identification of buried objects through microwave sensors (the so-called ground-penetrating radar) [199], the detection of underwater objects through acoustic signals and sonar [87], etc. Indeed, all of these scenarios critically involve multiple challenging problems of applied mathematics with the goal of formalizing and addressing recognition, classification, detection, tracking, estimation, inversion, and optimization tasks.

In this book, remote sensing is always intended as *remote sensing for Earth observation (EO)* [98, 169]. This means that the target, object, or entity of interest consists of a portion of the Earth surface or is located on the Earth surface, and that the information to be extracted is generally related to environmental applications. The sensor is on-board a satellite (*spaceborne sensor* or *satellite sensor*) or an aerial platform (*airborne sensor*) such as an aircraft, an unmanned aerial vehicle (UAV, also commonly called drone), or a balloon. The interaction between the observed area and the sensor occurs through electromagnetic waves.

EO technologies and methodologies have been acquiring a growing interest for the past few decades because they provide repetitive geographical area coverage and an increasing capability to observe the Earth surface at the desired spatial scale. Owing to the numerous space missions for EO, which have been deployed by national and international space agencies and by private industries, to the availability of airborne image acquisitions at national institutions and service companies, and to the currently growing interest of airborne lightweight platforms (i.e., the aforementioned UAVs), remote sensing currently provides a huge potential for environmental monitoring at global, regional, and local levels [98]. Remote sensing imagery represents a valuable source of information for a variety of applications, including but not restricted to vegetation-resource management and ecology (e.g., precision farming and forest inventory), urban planning (e.g., cadastral mapping and urban sprawl mitigation), oceanography (e.g., water-quality assessment and ocean current studies), hydrology (e.g., ice, snow, and drought monitoring), geology (e.g., stratigraphy studies), geophysics (e.g., crustal-dynamic monitoring and Earth magnetic field studies), renewable energy (e.g., biomass, irradiation, or wind speed resource assessment),

meteorology, climate change studies, defense, and security [91, 98]. EO also provides critical information in the framework of natural disasters (e.g., forest fires, floods, landslides, and earthquakes), both for prevention purposes (i.e., vulnerability assessment) and as a support to crisis management and post-crisis damage assessment [71].

This book addresses the broad field of the mathematical models and methods for the processing and the analysis of remote sensing images. Both tutorial aspects and the latest advances are discussed. Specifically, the objective of this introductory chapter is twofold. First, a summary of the fundamentals of remote sensing is presented to recall general concepts and basic terminology that will be used in the whole book (see Sect. 1.2). This discussion will be concise on purpose: More details on the major categories of two-dimensional remote sensing images will be provided in later chapters. Furthermore, the interested reader can find in the literature excellent textbooks on the fundamentals of remote sensing [29, 94, 117, 118, 169, 179, 180, 191, 207], on the physical bases and the instrumentation [45, 60, 61, 167, 176, 187, 199, 208, 209], on the applications [14, 42, 61, 78, 98, 106, 146, 164, 172], and on data processing and analysis [31, 34, 105, 151]. Second, an overview of the main families of mathematical models and methods for two-dimensional remote sensing image analysis is presented to provide the reader with a broad view of this scientific area before the other chapters focus on individual classes of models and algorithms (see Sect. 1.3).

1.2 Basics of Remote Sensing Imagery

1.2.1 The Notion of Remote Sensing Image

The data collected by EO sensors that will be considered in this book are formatted as two-dimensional (2D) digital images of a given geographic area and are named *remote sensing images* or *EO images* [169]. The goal of remote sensing image analysis methods is the extraction of geospatial information of interest from these images. Indeed, what is "of interest" (or what is not) is an application-dependent dilemma. It includes but is not limited to the mapping of land cover, land use, and of geophysical or biophysical properties of the observed surface, statically at a single time or dynamically along a temporal series.

It is well known that a *digital image* is a 2D table of points, named *pixels* (abbreviation of "picture elements"), which are associated with one or more discrete values (*pixel intensities*) [90, 203]. Given a digital image collected by a certain sensor, the meaning of the pixel intensities in terms of *measurements of physical quantities* substantially depends on the sensor itself [60]. The main physical quantities of interest for 2D remote sensing will be summarized in Sect. 1.2.3 and discussed in depth in later chapters.

From a signal-processing viewpoint, it is customary to consider an image as a realization of a 2D *stochastic process* (also often named *random field*) defined on the discrete lattice of the pixel grid [100]. This perspective is especially convenient whenever probabilistic and statistical modeling are necessary, a frequent situation when processing and analysis tasks have to be addressed [39, 163, 198]. Strictly speaking, remote sensing data are always discrete because they originate from a digitization process while they are collected by the sensor. Nevertheless, models and processing methods conveniently involve both discrete and continuous random processes. Numerous examples will appear throughout the book.

Scalar-valued and vector-valued random fields are used in the cases of a unique intensity or of multiple intensities associated with each pixel, respectively. From a computational standpoint, a scalar-valued image is actually a rectangular *matrix* whose numbers of rows and columns correspond to the height and width of the image, respectively. Analogously, a vector-valued image can be pictured as a *data cube*, whose sizes correspond to the width, the height, and the number of components of the vector-valued pixel intensities [31]. Such a data cube can be indirectly associated with a third-order tensor.

These three viewpoints on how to formalize what a remote sensing image is—a collection of physical measurements, a realization of a stochastic process, and a matrix or data cube—will be used consistently and interchangeably throughout the book.

1.2.2 Platforms for Remote Sensing

The acquisition of EO data is performed by one or more *sensors* that collect images on the observed scene and operate from one or more *platforms*. Both airborne and spaceborne platforms are used for remote sensing.

Major examples of *airborne* platforms include aircrafts, balloons, and UAVs [169]. Aircrafts that can be equipped with EO sensors are usually available at national or international organizations (e.g., military authorities) and at specialized companies. Balloons are overall less frequently employed for EO, but their use is steadily growing and may probably increase in the future. UAVs have been getting increasingly popular lately because of their low cost although they usually exhibit limitations on the maximum weight of the sensors they can carry. With both aircrafts and UAVs, altitude and orientation affect the geometry of the image, and acquisitions occur through *ad hoc* flights. Helicopters are also sometimes used for airborne remote sensing especially for specific applications. Processing examples using images from airborne platforms will be presented in Chaps. 2, 4, 5, 6 and 10.

A *spaceborne* platform is generally an artificial satellite orbiting around the Earth [169]. Exceptions include missions, such as the Spaceborne Imaging Radar-C / X-band Synthetic Aperture Radar (SIR-C/X-SAR) and the Shuttle Radar Topography

Mission (SRTM), in which sensors were put on-board NASA Space Shuttles [85]. Using the language of satellite missions, in a spaceborne EO system, the satellite represents the *space segment* while the *ground segment* is the infrastructure on the Earth surface that receives, validates, and pre-processes the acquired data. The space segment of an EO mission includes the *mission payload*, i.e., the sensor(s) that the satellite is designed to carry, and all necessary infrastructures for power, orbit management, on-board pre-processing, recording, and transmission to the ground segment [130].

It is well known that the path of a satellite along its orbit is an ellipse. The plane that includes this ellipse is named orbital plane. The orbits used for EO are geostationary or near polar. A *geostationary orbit* (or geosynchronous equatorial orbit, GEO) is circular, its orbital plane is the plane of the Earth Equator, and the orbiting period around the Earth is 24 hours [180]. Therefore, a sensor on-board a geostationary platform (geostationary sensor) always observes the same portion of the Earth surface. Simple calculations based on Newton's gravitation law imply that the altitude of a geostationary orbit is approximately 36000 km above the Earth Equator (for comparison purposes, recall that the mean Earth radius is estimated as 6371 km). Weather satellites (e.g., Meteosat Second Generation, MSG) are most often geostationary.

In the case of a *near polar orbit*, the angle between the orbital plane and the plane of the Equator is slightly larger than 90° (approximately 95–100°). Altitude is generally 400–1000 km—a range that is included within the broader family of low Earth orbits (LEO) [180]. Owing to the combination of the motion of the satellite around the Earth and of the rotation of the Earth itself on its axis, a sensor on-board a near polar satellite (near polar sensor) collects data over almost all the Earth surface. The projection of the satellite path on the Earth surface is usually named satellite ground track.

A special—and very often used—case of near polar orbit is the *Sun-synchronous orbit* (also known as heliosynchronous orbit). In this case, the angle between the orbital plane and the segment joining the centers of the Earth and the Sun is nearly constant in time [180]. Therefore, a sensor on-board a Sun-synchronous satellite (Sun-synchronous sensor) observes a given ground area at approximately the same time of the day on each consecutive overpass. This contributes to minimizing the differences in Sun illumination conditions across different observation times, an important property for the passive sensors that will be recalled in the next section. Examples of processing results using satellite images will be shown in almost all the chapters of the book.

1.2.3 Acquisition of Remote Sensing Images

Remote sensors for EO applications can be broadly categorized into two classes, i.e., passive sensors and active sensors. The latter transmit a signal toward the considered

surface and receive the resulting "echo" return. The former do not transmit any signal and directly receive the radiation incoming from the considered surface.

1.2.3.1 Passive Sensors for EO

A *passive sensor* receives the electromagnetic radiation that comes from the considered portion of the Earth surface either because it originates from the *reflection* of incident solar radiation or because it is *spontaneously emitted* by the surface itself.

Details on passive sensors and on the physical meaning of the quantities they collect can be found in Chaps. 2 and 3 and in textbooks such as [29, 167, 180]. Here, we only recall that the physical quantity measured by a passive EO sensor is the *spectral radiance* (or specific intensity). It is a radiometric quantity that characterizes the distribution of radiation in space, it represents the power per unit wavelength that travels in a unitary solid angle centered on a given direction through a unitary surface, and it is measured in $[\mathrm{W} \cdot \mathrm{m}^{-2} \cdot \mathrm{sr}^{-1} \cdot \mu\mathrm{m}^{-1}]$ [180, 194].

In the case of radiation in the visible portion of the electromagnetic spectrum (i.e., with wavelength between approximately 0.4 and 0.7 μm), in the near infrared range (NIR, 0.7–1.1 μm), and in the short-wave infrared range (SWIR, 1.1–1.35 μm, 1.4–1.8 μm, and 2–2.5 μm), spontaneous thermal emission from the Earth surface is negligible as compared to reflected solar radiation. Therefore, the spectral radiance received by a passive sensor operating in these ranges depends on the reflective properties of the observed surface, i.e., the *reflectance* [adimensional] and the *bidirectional reflectance distribution function* $[\mathrm{sr}^{-1}]$ that will be discussed in detail in Chap. 3 [105, 180].

Vice versa, in the case of radiation in the thermal infrared (TIR, also known as long-wave infrared, LWIR) portion of the spectrum (i.e., approximately 8–9.5 μm and 10–14 μm), reflected solar radiation is negligible as compared to Earth thermal emission [103, 180]. Therefore, the received radiance depends on the properties of the observed surface that characterize its capability to spontaneously emit radiation: Because of the well-known Planck's law, these quantities include the surface *temperature* [K] and *emittance* [adimensional] [105, 180]. In the intermediate case of mid-wave infrared radiation (MWIR, i.e., around 3–4 μm and 4.5–5 μm), reflection-based and emission-based contributions are comparable, and their reciprocal weights generally depend on all the aforementioned surface properties.

In all such cases, the spectral radiance that reaches the sensor first has to propagate through the portion of the atmosphere that is in between the surface and the sensor itself. In the case of reflected solar radiation, propagation through the atmosphere occurs twice, first downward from the direction of the Sun along the path from the

top of the atmosphere to the surface and then upward from the surface to the sensor. Propagation through the atmosphere, which is composed of a large number of particles, affects a spectral radiance field due to (i) the thermal emission of radiation by the atmosphere itself and (ii) the *extinction* of the propagated radiance field due to absorption (i.e., conversion of part of the energy associated with the radiation to heat) and to scattering from one propagation direction to another (i.e., redistribution of the energy associated with the radiation through different directions) [194]. These phenomena are quantitatively well described by the so-called *radiative transfer equation*, which is an integro-differential equation that can be explained in terms of conservation of energy [194] and of scattering in random media [143]. The solution is generally a complex problem for which specific numerical techniques have been developed [143, 188, 194].

A passive EO sensor is most often designed to be *multispectral*, i.e., it collects data simultaneously from multiple wavelength ranges, named *bands* or *channels* [105]. In particular, one speaks of a *hyperspectral* sensor if a large number (usually hundreds) of channels with narrow bandwidths are collected [129]. Vice versa, a sensor designed to acquire only one channel, which usually encompasses the whole visible (and possibly NIR) range, is named *panchromatic* [6].

Multispectral acquisition can be accomplished using prisms and optical filters, which split the incoming radiance into different wavelength ranges, or using separate cameras that operate in distinct wavelength ranges directly [180]. Figures 1.1 and 1.2 show a multispectral image acquired in 2004 by the IKONOS sensor over Metaponto, Italy, and several channels of a hyperspectral image collected on August 23, 1995, by the airborne HYDICE sensor over Washington DC, USA.[1] The former is composed of four channels, approximately corresponding to the blue, green, red, and NIR wavelength ranges. Examples of color composites, in which the R, G, and B components of a displayed color image [203] are associated with three of the available channels of the multispectral remote sensing images, are also shown in Fig. 1.1e, f. The hyperspectral image of Fig. 1.2 is composed of 210 channels across the visible, NIR, and SWIR ranges. One can note the strong impact of atmospheric extinction [194] in certain wavelength ranges. More generally, it is worth noting that, because of the aforementioned physical processes that lead to image formation, data collected by passive sensors are obviously affected by atmospheric (e.g., cloud cover) and Sun-illumination conditions.

[1]This hyperspectral image can be downloaded from the Purdue University Web site at https://engineering.purdue.edu/~biehl/MultiSpec/hyperspectral.html.

(a) Blue (b) Green (c) Red

(d) NIR (e) True color composite (f) False color composite

Fig. 1.1 Example of a multispectral image acquired by the passive IKONOS sensor over Metaponto, Italy (1250 × 1250 pixels): spectral channels corresponding to *blue* **(a)**, *green* **(b)**, *red* **(c)**, and NIR radiation **(d)**; the *true color* composite **(e)**, in which the R, G, and B components of the displayed image are associated with the *red*, *green*, and *blue* channels of the multispectral image, respectively; and a *false color* composite **(f)**, in which the R, G, and B components of the displayed image are associated with the NIR, *red*, and *green* channels of the multispectral image, respectively

1.2.3.2 Active Radar Sensors for EO

An *active sensor* transmits an electromagnetic pulse toward the considered portion of the Earth surface and receives the resulting "echo" signal. For the purpose of 2D remote sensing image acquisition, *microwave* signals are typically used, and the imaging system is based on a *radar* (RAdio Detection And Ranging) instrument [168, 199, 209].

A laser source can also be used for active remote sensing in a LiDAR (Light Detection And Ranging) instrument, also known as airborne laser scanning (ALS) or LaDAR (Laser Detection And Ranging) [68]. LiDAR is a most prominent technology for 3D mapping through remote sensing. As this book is focused on 2D remote sensing image analysis, LiDAR will not be discussed any further.

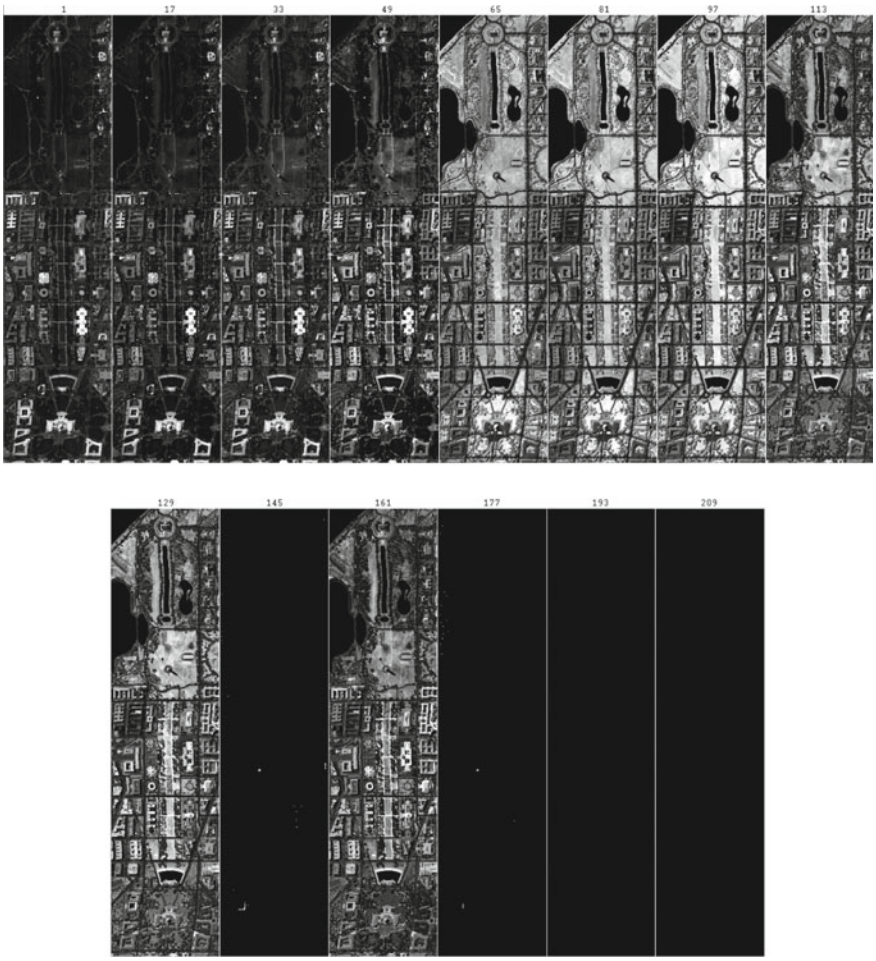


Fig. 1.2 Example of channels from a hyperspectral image taken by the airborne HYDICE sensor over Washington DC, USA (307 × 1280 pixels, 210 bands in the 0.4–2.4 μm wavelength range—visible, NIR, and SWIR). Very *dark* bands are obtained in wavelength ranges in which severe atmospheric extinction occurs

A detailed discussion of radar systems for remote sensing image acquisition can be found in Chap. 4. Here, we recall that a radar imager for EO periodically emits a short-duration microwave pulse that is irradiated by a directive antenna as an electromagnetic wave in space [12]. Part of the irradiated energy hits the considered surface that re-irradiates it in multiple directions, a phenomenon known as *scattering* [194]. The portion of the re-irradiated signal that is backscattered in the direction of the antenna is received by the antenna itself. In the application of radar to positioning, the backscattered signal can be used for detecting the presence of a given target object

(e.g., an aircraft), for measuring the distance of this target through the time taken by the pulse to reach the target and get back to the antenna and for estimating the speed of the target through the Doppler effect [199].

In the case of remote sensing image acquisition, the antenna is put on-board an airborne or spaceborne platform, and the goal is to use the backscattered signal to measure electromagnetic properties of the considered portion of the Earth surface in the microwave range. In the basic configuration of a single-frequency and single-polarization radar system for EO, the main property that is measured is the *backscattering coefficient* [adimensional], which is related to the average power of the return signal (see [199] and Chap. 4). It is affected by numerous factors, including the roughness of the surface, its moisture content if it is a soil area, the presence on the surface of 3D structures (e.g., buildings), the carrier frequency of the microwave pulse, and the radar polarization. Polarization properties will be discussed in Chap. 5.

Regarding the *carrier frequency* and the corresponding wavelength, we recall that, in general, the word "microwave" broadly refers to electromagnetic waves with frequency between 1 and 100 GHz, although precise definitions may vary [168]. Specifically using the IEEE Std 521 standard for radar frequency bands [1], we can mention the L-band (i.e., 1–2 GHz of carrier frequency or equivalently 15–30 cm of wavelength), the C-band (i.e., 4–8 GHz or 3.75–7.5 cm), and the X-band (i.e., 8–12 GHz or 2.5–3.75 cm) among the most common ranges for radar EO.

The *pulse signal* used by a radar for EO exhibits a narrowband spectrum in a neighborhood of the carrier frequency and is most usually a linearly frequency-modulated signal known as *chirp* [168]. This choice, together with appropriate filtering of the return signal, makes it possible to achieve high spatial resolution along the looking direction of the radar (named the *range direction*) [199].

The *synthetic aperture radar* (SAR) technique makes use of the motion of the platform along its path to simulate a long antenna, which, in turn, makes it possible to achieve high spatial resolution along the flight direction (named the *azimuth direction*) from both airborne and spaceborne platforms [66, 135, 199].

Details on SAR processing and its implications can be found in Chap. 4. Four examples of SAR images are shown in Fig. 1.3. Images acquired by the C-band Sentinel-1A sensor over Marseille, France, on April 14, 2017, by the C-band RADARSAT-2 sensor over Port-au-Prince, Haiti, on December 8, 2011, by the X-band COSMO-SkyMed mission over the aforementioned area of Metaponto, Italy, on March 3, 2011, and by the L-band PALSAR-2 sensor on-board the ALOS-2 satellite over Panama on September 1, 2015 are shown in Fig. 1.3a, b, c, and d, respectively.

It is worth noting that a radar system for EO operates regardless of Sun illumination, because it makes use of its own source of transmitted energy, and that the resulting data are almost insensitive to cloud cover and atmospheric conditions [168]. Therefore, unlike passive instruments, radar sensors for EO provide day-and-night

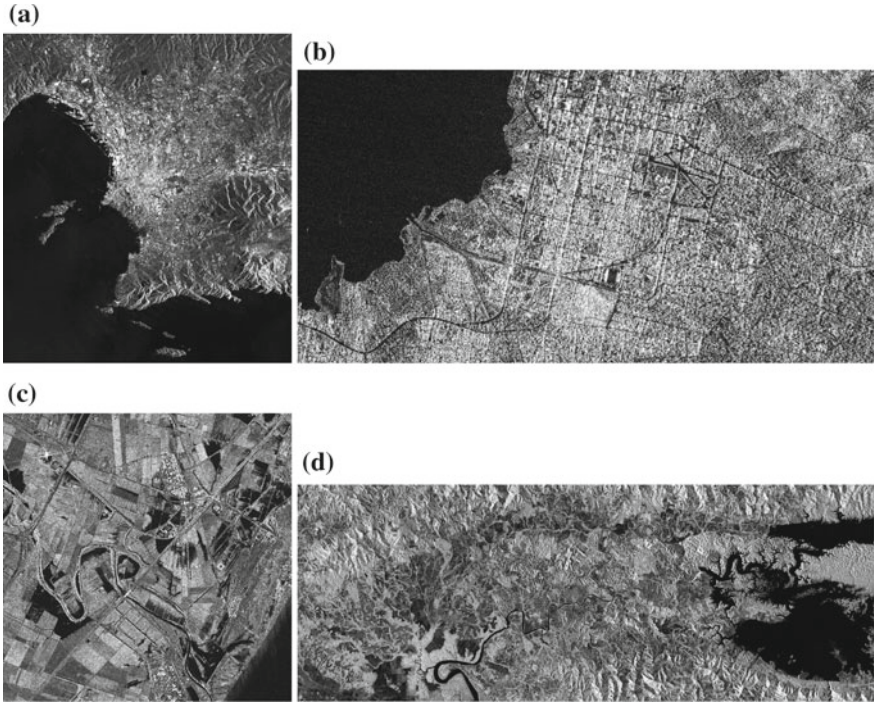


Fig. 1.3 Examples of SAR images: **a** Sentinel-1A image acquired over Marseille, France (2100 × 2400 pixels; ©ESA); **b** RADARSAT-2 image acquired over Port-au-Prince, Haiti (1536 × 781 pixels; ©MDA); **c** COSMO-SkyMed (COnstellation of small Satellites for Mediterranean basin Observation) image acquired over the same area of Fig. 1.1 shortly after a flood (2000 × 2000 pixels; ©ASI); **d** PALSAR-2 (Phased Array type L-band Synthetic Aperture Radar 2, on-board the Advanced Land Observing Satellite 2, ALOS-2) image of a vegetated area and of Lake Bayano in Panama (9228 × 3471 pixels; ©JAXA, distributed by PASCO CORPORATION)

and all-weather acquisition capability. Indeed, the complementarity between the properties of passive multispectral and active radar imagery will be discussed in more detail in Chap. 7.

In addition to the basic single-frequency and single-polarization mode, other configurations of radar EO also exist. *Polarimetric SAR* (PolSAR), which will be discussed in Chap. 5, collects (usually complex-valued) measurements associated with multiple polarizations simultaneously [41, 111]. *Interferometric SAR* (InSAR), which will be analyzed in Chap. 4, exploits measurements of the phase of the radar return (and not only of its power) to extract 3D information on the observed surface [66]. *Differential InSAR* (DInSAR) further extends InSAR to map slow movements of the surface (e.g., due to seismic phenomena) [135]. *SAR tomography* uses measurements taken from different altitudes (e.g., different orbits) to characterize the vertical structures of the targets [168]. *Multifrequency SAR* uses multiple antennas on-board the same platform to collect data at multiple carrier frequencies [85].

Bistatic SAR uses distinct antennas for transmittance and reception to investigate the scattering behavior in multiple directions [199].

1.2.4 The Notions of Resolution

The word "resolution" has already been used a couple of times in the previous sections and intuitively refers to the precision with which a given instrument captures details on the observed surface. More precisely, the specific meaning of "resolution" depends on the domain it is referred to.

First, *spatial resolution* is the size of the smallest spatial detail that can be distinguished in a remote sensing image [169]. It is obviously related to the size of the ground area associated with a pixel, although it can be—and generally is—at least slightly coarser than this size because of blurring effects within the acquisition chain. In the case of a passive sensor, the spatial resolution depends on the sensor optics and on the altitude of the platform [180]. In the case of a SAR instrument, it is possible to prove that the spatial resolution resulting from chirp processing, and SAR technology is independent on the platform altitude [168, 199]. The spatial resolutions of current satellite sensors for civil applications are approximately a few kilometers in the case of geostationary sensors (e.g., 3 km for the TIR bands of the SEVIRI sensor on-board MSG [178]), a few tens of meters in the case of moderate resolution sensors (e.g., the Landsat series of satellite missions [120]), and up to 30 cm–1 m with recent very high resolution (VHR) near polar sensors (e.g., WorldView-2 and -3 [148], Pléiades [22], COSMO-SkyMed [205], and TerraSAR-X [213]; see Chaps. 3 and 4). Spatial resolutions up to a few centimeters can usually be obtained using airborne acquisitions. For example, Fig. 1.4 displays portions of six remote sensing images with the same size in pixels (400 × 400 pixels) and with very different spatial resolutions (3 km, 500 m, 30 m, 10 m, 2 m, and 5 cm). The difference in the spatial details that can be appreciated in these images is visually evident.

The *temporal resolution* of a spaceborne sensor is the frequency with which a given ground area is repetitively observed. It is generally expressed in terms of the *revisit time*, i.e., the time between two consecutive satellite overpasses [169]. Typical values range from a few tens of minutes for geostationary sensors (e.g., approximately 15 min for MSG [178]) to a few days or weeks for near polar sensors (e.g., one day for the Moderate Resolution Imaging Spectroradiometer, MODIS [95], and 16 days for Landsat-8 [120]; see also Chap. 9). The use of multiple satellites in a constellation favors shorter revisit time (e.g., up to 12 hours for the COSMO-SkyMed constellation composed of four satellites [205]). It is also worth noting that current near polar sensors often exhibit a pointing functionality, i.e., their observation directions can be steered upon agreement with the agency or company in charge of mission operations. This allows more frequent observations to be obtained on a given area but could sometimes make revisit times no more periodical and less predictable.

(a) 3 km (b) 500 m (c) 30 m

(d) 10 m (e) 2 m (f) 5 cm

Fig. 1.4 Color composites of channels from six multispectral images of size equal to 400 × 400 pixels: **a** MSG Spinning Enhanced Visible and Infrared Imager (SEVIRI) image of South Africa (©ESA); **b** Sentinel-3 Sea and Land Surface Temperature Radiometer (SLSTR) image of the south coast of France (Occitanie and Provence-Alpes-Cote d'Azur regions; ©ESA); **c** Landsat-8 Operational Land Imager (OLI) image of Nice, France (©USGS); **d** Sentinel-2 image of Genoa, Italy (©ESA); and **e** Pléiades image of Venice, Italy (©CNES distribution Airbus DS); and **f** image collected by an airborne sensor over Zeebruges, Belgium ([grss_dfc_2015] data set [2]). The spatial resolution is indicated below each image

The *spectral resolution* is associated with passive multispectral sensors and is the precision with which the incoming radiation is sampled along the electromagnetic spectrum. It is usually expressed in terms of the number of channels of the sensor and of the widths of the corresponding wavelength ranges. Current sensors for civil applications range from a few bands of moderate width (70–100 nm each) to the case of hyperspectral sensors with a few hundreds narrow bands (2–10 nm each; see Fig. 1.2 and Chap. 2). For reasons associated with signal-to-noise ratio, a trade-off usually exists between the spectral and the spatial resolutions of a given passive sensor [105]. Therefore, several current satellite passive systems carry both a multispectral sensor, with several bands in the visible and NIR range, and an additional panchromatic sensor, which has obviously poorer spectral resolution but achieves finer spatial resolution than the multispectral bands [105, 148] (see also Chaps. 6 and 7). For example, Fig. 1.5 shows the panchromatic band at 1-m spatial resolution and a color composite of three spectral channels at 4-m spatial resolution collected simultaneously by IKONOS over Alessandria, Italy.

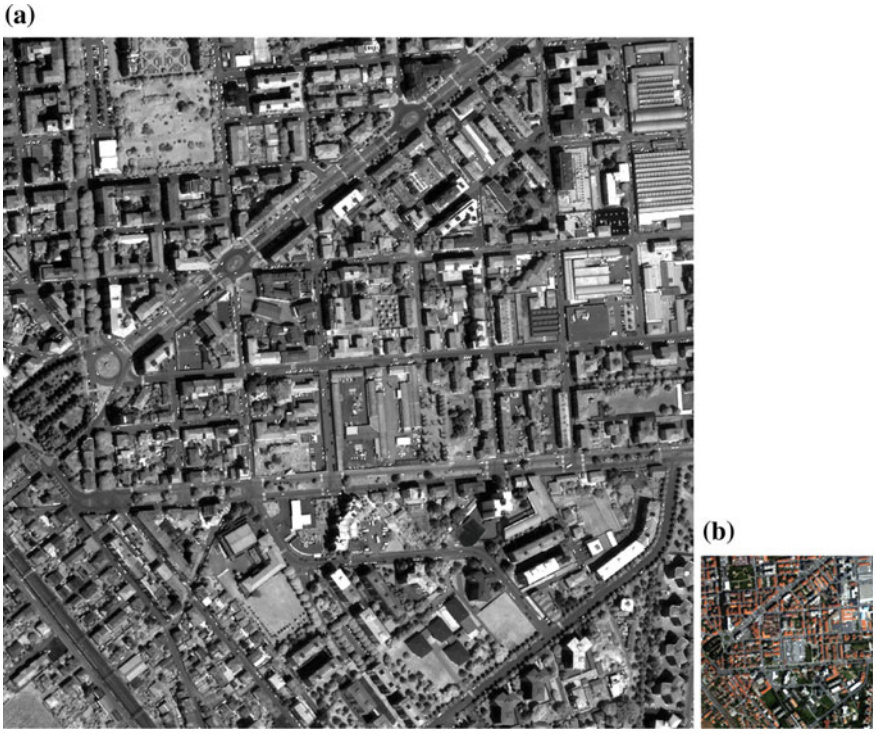


Fig. 1.5 Portion of an IKONOS image collected over Alessandria, Italy: **a** panchromatic channel at 1-m spatial resolution; **b** color composite of the *red*, *green*, and *blue* multispectral channels at 4-m spatial resolution

Finally, *radiometric resolution* is related to the precision with which differences in the considered physical quantities can be appreciated and measured in the recorded image [169]. It is related to the signal-to-noise ratio of the sensor [105] and to the digitization process that is included in the acquisition chain. The intensity of a pixel in a digital image (or in each band of a multispectral digital image) is encoded with a finite number of bits, which correspond to a finite number of levels in a predefined discrete set (named *quantization levels* in the signal processing literature and often *digital numbers* in the remote sensing literature). The radiometric resolution of a sensor is generally expressed in terms of the number of bits that are used to encode each quantized intensity and are associated with each pixel (number of bits per pixel, bpp, sometimes also named bit depth). Typical values range from 8 bpp (256 levels) to 12 bpp (4096 levels) and 16 bpp (65536 levels).

1.3 Mathematical Modeling for Remote Sensing Image Analysis

1.3.1 General Comments

The main functional stages of a general system for the interpretation of remote sensing imagery can be described by the block diagram in Fig. 1.6 [105, 118]. After the acquisition phase, whose main ideas were recalled in the previous sections, *pre-processing* operations are generally necessary before the data are used to extract thematic, biophysical, or geophysical information. They may include calibration, registration, as well as radiometric and geometric correction. Georeferencing (also known as geocoding) is also necessary to relate the pixel coordinates in an image to geographical coordinates in a cartographic system. Overviews of remote sensing image pre-processing are reported in [34, 129, 151, 169], and a comprehensive treatment of registration methodologies can be found in [109]. Calibration aspects associated with multispectral sensors will also be discussed in Chap. 3.

The resulting pre-processed imagery can be fed to the *analysis* stage, which aims at extracting, from the image data, the information of interest to a given end user. This information typically consists of a set of maps or of further transformed images associated with the considered geographical area. The meaning of this output varies substantially as a function of the objective of the analysis task and of the type of input imagery. Customary examples include but are not restricted to the following cases:

- In an output *classification* map, each pixel is assigned to one of a finite number of classes corresponding, for example, to land-cover or land-use categories [18, 34, 50, 128] (see Chaps. 2 and 10).
- In the case of an output *regression* map, the input imagery is used to retrieve indirect measures of biophysical or geophysical parameters of the observed portion of the Earth surface (e.g., land surface temperature, chlorophyll concentration in sea water, or wind speed over ocean areas) [4, 103] (see Chap. 10).
- In the case of a *segmentation* map, the input image is partitioned into a finite set of homogeneous, often connected, regions (also named segments or superpixels) [39, 50];

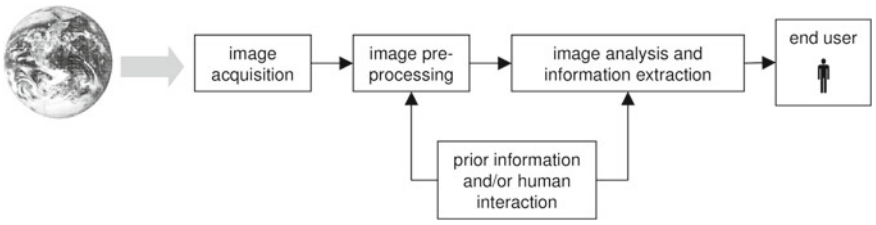


Fig. 1.6 Block diagram of a generic system for the interpretation of remote sensing imagery

- In an *object detection* (possibly target detection) result, the presence and locations of specific objects or structures in the imaged scene are determined [44, 53, 129, 162, 184].
- If input images taken at different times are available, a *change detection* result characterizes whether and how the ground area related to each pixel has changed [13, 34, 104] (see Chaps. 8 and 9).
- If input hyperspectral imagery is used, *unmixing* results, in which the observed spectral behaviors are decomposed into a collection of "pure" components, can be derived as well [162] (see Chap. 2).
- If the input imagery originates from multiple data sources, the output is often intended as a *data fusion* result [3, 169] (see Chaps. 6 and 7).
- If an entire database of input images is available, then *data mining* results, which generally aim at discovering knowledge in the database, can be considered [112] (see Chaps. 3 and 9).

Although both the pre-processing and the analysis stages are usually meant to be as automatic as possible, human interaction can generally be involved, for example, to incorporate prior knowledge about the considered area into the processing chain.

The necessary methodological basis for developing effective and computationally efficient solutions to problems of 2D remote sensing image analysis derives from *pattern recognition*, *machine learning*, and *image processing* [31, 38, 105]. These disciplines are rooted in or related to numerous areas of applied mathematics including probability theory, statistics, random process theory, graph theory, functional analysis, numerical analysis, operations research, and information theory.

With regard to terminology, pattern recognition and machine learning, although they derived from different fields—i.e., signal processing and computer science, respectively—can essentially be considered as two facets of the same scientific discipline [21]. For the fundamentals of pattern recognition and image processing, we refer the reader to well-known textbooks such as [21, 69, 90, 160, 203]. Nevertheless, in all chapters of the present book, not only advanced mathematical models and methods based on these disciplines will be presented but also basic concepts will be reviewed.

In the next subsections, major families of mathematical models and methods, which are prominent for 2D remote sensing image analysis and are rooted in pattern recognition and image processing, are recalled. Computational techniques and case studies based on these families of methods will appear in all other chapters of the book.

1.3.2 Mathematical Models for Image Data Representation

While classical statistical pattern recognition methods often work under the assumption of independent and identically distributed (i.i.d.) samples [21, 69], the analysis of remote sensing images normally requires models that can represent dependencies among the data. For example, relationships in the spatial domain are of primary importance, especially when VHR data are concerned, because of the strong correlation among neighboring pixels, the presence of textures, or the dependencies among regions and objects in the scene [50]. Furthermore, in a remote sensing image, each pixel is most often associated with a vector of pixel intensities, because multiple measurements are simultaneously collected by multispectral and polarimetric SAR sensors and because additional features can also be computed from the measured data. Therefore, multivariate dependency models for data in multidimensional vector spaces, possibly even supporting the case of high dimensionality (e.g., with hyperspectral images), are necessary [31, 105]. Image representation methods are also required for data mining purposes and to move from statistical to more semantic-oriented descriptions of the content of an image [112]. Within this framework, mathematical models and methods that provide effective tools for image data representation are especially relevant.

A major family of methods for data representation is based on *graph theory* [55]. A graph is a natural model for a set of dependent entities, which are identified with the vertices (or nodes) of the graph and are linked with edges as a function of the dependency structure to be described. A special case, which is involved in numerous remote sensing image analysis techniques, occurs when the graph is a *tree*. In this case, the graphical model intrinsically describes *hierarchical* relationships from the root to the leaves of the tree. The meaning of the nodes and edges of a graphical model vary according to the kind of dependencies that are being represented (spatial, multiscale, semantic, etc.). Special cases of image data representation models that are generally based on or related to graphs include *probabilistic graphical models* (see next Section) [99], multiscale *wavelet decompositions* [127], tree models associated with *mathematical morphology* [93, 185], and *manifold learning* approaches [125].

For instance, the nodes may be pixels, regularly or arbitrarily shaped regions, objects, etc., and the edges may indicate spatial neighborhood relationships, temporal correlations, multiscale dependencies, semantic interconnections, etc. Graphical models have been applied in numerous remote sensing data analysis methods to characterize spatial, region-based, temporal, multiscale, or multiresolution information among others [5, 49, 74, 84, 122, 197] (see also Chaps. 2, 3, 4, 6, and 7). Further generalizations that make use of hypergraphs, in which edges are replaced by subsets of more than two vertices to capture more general relationships, have also been considered [92].

In particular, hierarchical tree-based image representations are naturally obtained when multiscale image transforms are applied. Rooted in harmonic analysis [171], *wavelets* allow transformed images that capture information at different spatial scales to be computed and multiscale image representations to be constructed. They originate from the theory of multiresolution signal decompositions of continuous-time signals in the L^2 space [126] and generalize the Fourier transform by mapping to a transformed domain in which temporal localization and scale-dependent behaviors are studied [141]. In the discrete case of digital images, they are formalized through the combination of appropriate low-pass and high-pass filters and downsampling operations on the pixel lattice, which result in a tree representation of the input image (see also Chaps. 7 and 8) [127]. Many approaches that use these hierarchical representations for image processing, analysis, and fusion tasks have been proposed in the remote sensing literature [6, 9, 27, 84, 109, 211]. Analogously, *time-frequency analysis* provides linear operators that allow analyzing a continuous-time signal simultaneously in the time and frequency domains [127, 141]. The short-time Fourier transform, the Gabor transform, and the Wigner distribution function are major examples and have been applied in remote sensing as well [59, 70, 114] (see also Chap. 5).

Furthermore, tree representations can also be related to *mathematical morphology* [183]. In this framework, which generalizes Minkowski's set theory, a powerful set of nonlinear operators (named morphological operators), which favor sensitivity to shape and size properties and contribute to characterizing objects in the imaged scene, is introduced. They include erosion, dilation, opening and closing operators, geodesic transformations, and operators by reconstruction among others [183, 185]. They can also be formulated so that they provide a multiscale decomposition of the imaged scene in terms of *morphological profiles* (see Chaps. 6 and 7) [159]. A recent generalization is also given by *attribute filters*, whose result depends on the evaluation of a predicate on connected image components, and which provide multiscale representation capability (*attribute profiles*) [18, 163, 185] in association with component trees and trees of shapes (see Chaps. 3 and 7) [35, 76, 93].

Image representation methods based on *manifold learning* are also especially relevant for high-dimensional data [125]. In the language of topology, an m-dimensional manifold is a subset of a topological space that is locally homeomorphic to an open subset of an m-dimensional Euclidean space [110]. Classical examples are simple regular curves and surfaces in a 3D space. In the case of a remote sensing image composed of a large number n of bands, the n-dimensional data samples most typically belong, up to noise, to some lower-dimensional subset that can be identified with an m-dimensional manifold ($m < n$). For example, in the case of a hyperspectral image, an explanation is that in general, the relationships among the bands are highly nonlinear because of the complex interactions among the spectral radiance field, the Earth surface, and the atmosphere [163]. Manifold learning methods for remote sensing image analysis aim at capturing the manifold structure by nonlinearly projecting the high-dimensional data to a lower-dimensional space in such a way to preserve the local topology [11]. A graph such that the nodes are the data points and the edges are based on local distances (e.g., through the k nearest neighbors

algorithm) is typically used as a discrete sampling of the manifold. Then, *spectral graph theory*, which characterizes the properties of a graph in terms of suitable matrices and of their eigenvalues and eigenvectors, can be used to capture the embedding of the manifold into the n-dimensional space (see Chap. 10) [121, 125, 195].

We also recall important approaches to data representation based on *sparse models* and *compressive sensing* (or *compressed sensing*). The key idea is that most signals actually exhibit a substantial degree of redundancy, or equivalently, they can be mapped to some transformed domain in which they are sparse. Sparse models and methods aim at representing data as sparse linear combinations of elements (named *atoms*) drawn from a finite overcomplete *dictionary* [158].

From a signal processing perspective, the redundancy property is intuitively consistent with common properties (extensively used for compression purposes) of audio, image, and video signals and of Fourier-like and wavelet transforms [127, 160]. More precisely, it has been proven that given suitable knowledge of the sparsity of a signal, substantially fewer samples than those implied by the Nyquist-Shannon sampling theorem can actually be enough to reconstruct the signal itself [33, 56]. In the specific case of remote sensing images, redundancy is intrinsically favored by the aforementioned high resolutions in the spatial and/or spectral domains. Sparse and compressive sensing models have been found useful in remote sensing image analysis problems with multispectral [6, 72, 119], SAR [9, 212], and especially hyperspectral data [20, 147, 196]. Comments on their computational aspects are reported in Sect. 1.3.5 and in Chap. 9.

1.3.3 Probabilistic Modeling and Bayesian Methods to Learn from Image Data

As mentioned in Sect. 1.2.1, it is customary to think of a remote sensing image as a realization of a 2D random field. From a modeling perspective, this is consistent with the approach used in many (1D, 2D, and nD) fields of signal processing. More specifically, remote sensing data should be considered as intrinsically characterized by measurement uncertainty. For instance, the spectral radiance measured by a passive sensor is affected by the state of the observed surface, the individual objects included in the ground region corresponding to each pixel, the atmospheric effects, and the instrumentation noise. All such contributions cannot be characterized deterministically and require stochastic modeling. Similar comments obviously hold with regard to SAR imagery as well [65].

Many approaches to remote sensing image analysis are formulated in terms of *probabilistic modeling* through stochastic processes [100]. From this perspective, a natural class of mathematical techniques to formalize learning and recognition tasks includes *Bayesian* methods, ranging from the well-known *decision theory* [200] to more recent approaches such as *Gaussian process regression* [166] and *marked point processes* [54], and often involving *probability density estimation* [21] problems and *estimation theory* concepts [200].

A consolidated approach in statistical pattern recognition, Bayesian learning is rooted in statistics and probability theory and is at the basis of numerous methods of image classification [105, 151], change detection [28, 79], target and anomaly detection [136, 147], unmixing [20], data fusion [79, 186], and bio/geophysical parameter regression [32] (see Chaps. 4, 5, 7, 9, and 10). In particular, the Bayesian decision theory is a general-purpose mathematical model to formalize decision problems in a probabilistic framework; classification and detection problems are stated as special cases of decision. The well-known maximum *a-posteriori* (MAP), maximum likelihood (ML), minimum risk, minimax, Neyman-Pearson, and marginal posterior mode (MPM) rules are formalized within this theory and are endowed with remarkable properties of optimality (e.g., MAP and MPM minimize suitable probabilistic functionals related to the classification errors, given a model for the data statistics; see also Chap. 7) [132, 200].

Indeed, the probabilistic standpoint used by the Bayesian approach implies that data distribution is assumed to be perfectly known. In the application to image analysis, it is usually not and has to be estimated from the available image data [105]. This issue relates Bayesian approaches to *estimation theory* meant as the class of mathematical methods for estimating the parameters of a given family of distributions from a set of samples [200] and more generally to *probability density estimation* [21, 69]. Specifically, when a *parametric model* (e.g., Gaussian, Gamma, Fisher, or Gaussian mixture) is postulated for the unknown distribution of a random vector, the estimation theory provides both theoretical bounds on the estimation performances (e.g., the Cramér-Rao bound) and computational parameter estimation algorithms. The well-known ML, MAP, minimum mean square error, method of moments [200], and expectation-maximization (EM) [52] approaches have been applied, along with several extensions (e.g., the stochastic EM [36]), within numerous processing and analysis techniques for multispectral, hyperspectral, and SAR data (see also Chaps. 7 and 9) [9, 23, 51, 67, 72, 105, 115, 182, 193, 202]. The method of log-cumulants, which extends the well-known method of moments by using Mellin transforms (instead of the usual Laplace transforms) to define a new kind of moment generating functions, has even been initially developed for parameter estimation tasks in the SAR field (see Chaps. 4, 5, and 7) [8, 101, 149, 198].

When a *nonparametric density estimation method* is used, no prescribed model is assumed for the unknown distribution, and a density estimate is computed using the available samples directly. Well-known approaches include the Parzen window

algorithm [21], the k nearest neighbors method [69], and truncated orthogonal polynomial expansions (e.g., the Gram-Charlier and Edgeworth expansions [97]). While they generally enhance modeling flexibility as compared to parametric models, they are also more prone to overfitting and may include, in turn, additional internal parameters (sometimes named hyperparameters) to be optimized as well. Nonparametric density estimators have been used in remote sensing not only for classification [49] or change detection [23] but also in conjunction with information-theoretic functionals [3, 89].

In general terms, *information theory* is a well-known mathematical formalization for source and channel coding in communication systems and allows the information conveyed by a given random source to be quantitatively measured on a probabilistic basis [10]. Methodological results from information theory, which have been used for remote sensing data modeling and mining [3, 37, 89, 139], also allow quantifying the distance between probability distributions as well as the mutual information, the dependency, and the complexity associated with multiple random sources (see Chaps. 8 and 9) [10].

Among the Bayesian nonparametric methods, *Gaussian process regression* has recently been found attractive in various applications to biophysical and geophysical parameter retrieval [32, 156]. This approach models the relationship between the quantity to be estimated and the input data as a multidimensional Gaussian stochastic process [166] whose autocovariance function can be defined to incorporate the desired spatial, spatiotemporal, or stationarity behavior [32]. Regression is formulated according to a Bayesian MAP criterion, while estimation-theory methods are applied to determine the hyperparameters of the method (see Chap. 10).

The use of Bayesian formulations together with graph models for image data representation leads to *probabilistic graphical models* [99]. In this framework, Bayesian rules are formalized in accordance with the graph topology and generally in conjunction with suitable Markovianity assumptions. The resulting models have proven powerful and flexible for characterizing spatial, spatiotemporal, multiresolution, and multisensor information [86, 113, 115, 144, 175, 177, 186, 198]. Major examples are represented by *Markov random fields (MRF)* and *conditional random fields (CRF)*. An MRF characterizes, in terms of local relationships, the prior information associated with the latent (or hidden) random field to be estimated (e.g., the field of the class memberships of the pixels in a classification problem; see also Chaps. 4 and 7) [73, 96, 116, 132]. In the case of a CRF, Markovianity is assumed for the posterior distribution directly [190]. In both cases, Bayesian rules are expressed as the minimization of suitable *energy functions*. Other examples include pairwise and triplet Markov models, which are especially useful to characterize non-stationary image behaviors [161], and Bayesian networks (or belief networks), which make use of directed acyclic graphs to relate a set of random variables [48]. From a machine learning perspective, these models fall under the family of *structured output learning*, which includes techniques whose output is composed of mutually dependent samples [150]. From a mathematical viewpoint, they are at the crossroad among inferential statistics, graph theory, and combinatorial optimization (see also Sect. 1.3.5).

Finally, an even more general family of probabilistic models for complex spatial relationships is given by *marked point processes*. Originated from stochastic geometry, they are advanced and mathematically elegant models for populations of objects that can be randomly distributed across the imaged scene [54]. They have been found especially effective for the simultaneous detection of multiple objects (e.g., trees, cars, buildings, or boats) in VHR optical images [17, 44, 53, 152, 192]. In a marked point process model, a 2D Poisson process, which is meant to localize the objects in the image, is augmented with further random variables that parameterize geometrical properties of the objects [54]. Similar to the cases of MRF and CRF, methods based on marked point processes generally use Bayesian minimum-energy rules. However, unlike a probabilistic graphical model, a marked point process is not attached to a prescribed graph—a remarkable property when it is necessary to capture irregular and complex arrangements of objects whose number is originally unknown.

1.3.4 Non-Bayesian Methods for Learning from Image Data

Besides the approaches recalled in the previous section, image interpretation problems involving remote sensing data can also be formalized and successfully addressed by resorting to other families of methods, which do not make use of Bayesian formulations and sometimes do not even explicitly involve any probabilistic model for the data [21]. On the contrary, they build on top of concepts drawn from other areas of applied mathematics, such as Hilbert space theory, functional approximations, or set and logic theory. Among the main families of non-Bayesian learning methods, we recall the *neural*, *kernel*, *tree*, and *fuzzy* approaches.

Initially inspired by biological modeling, neural networks have evolved into a vast and powerful class of methods for classification and regression [21, 58]. They have recently become particularly prominent because of the outstanding performances provided by deep neural networks [80, 107] in computer vision applications.

> *Artificial neural networks* are characterized by (often massively) parallel architectures, which ensure adaptivity and nonparametric learning capability. The input–output relationship of the network is determined by a cascade of interconnected *layers*, each composed of a finite set of parallel nonlinear processing units named *neurons* [58]. This relationship can be used as a nonparametric *functional approximator* of a generic function and of a posterior probability distribution in the cases of regression and classification, respectively.

The architecture of a neural network can be related again to a graphical model, in which the number of layers is a measure of the "depth" of the network. Optimal values for the parameters that determine the input–output relationship are learned on the basis of input data (see also Sect. 1.3.5). From a mathematical viewpoint, it has been proven that, under mild assumptions, the subspace of the neural functional

approximators is dense in the Banach space of continuous functions on a compact set, i.e., neural networks can approximate arbitrary continuous functions uniformly on compact sets—a statement often named *universal approximation property* [47]. Similar results and bounds also hold in other metric spaces (e.g., L^p) [15]. Major examples, which differ in the network topology and learning processes and which have been applied to remote sensing image analysis [7, 43, 64, 154, 204], include the multilayer perceptron, radial basis function, Kohonen self-organizing maps, adaptive resonance theory, Hopfield, Boltzmann machine, associative memory, and recurrent networks among others [58].

While shallow networks, which are composed of only a few layers, were mostly used for some decades, *deep networks*, which include many layers in cascade, have lately been attracting strong attention in the learning and recognition community and in remote sensing as well [80, 107]. The intuitive rationale is that, across the numerous layers, the network generates representations of the input data at progressively more and more abstract levels, which are not only customized to a specific processing task but are also directly learned from a large data set. Although these concepts have been known methodologically since the 1980 s and 1990s, they became prominent when high-performance computing architectures (e.g., graphical processing units) [162] made it possible to operate with large networks and millions of input data samples in affordable times [107]. The resulting *deep learning* architectures turned out to be effective at capturing intricate structures in possibly high-dimensional data. With regard to the analysis of images or data cubes, a major case is represented by *convolutional neural networks (CNN)*, which were recently found successful in remote sensing image processing, classification, and mining [131, 134, 153, 170, 206]. They include case-specific layers that apply convolution operators and pooling procedures to capture spatial information and favor translation invariance [80]. The resulting number of network parameters is typically huge, which requires very large data sets for learning purposes. Case-specific procedures that combine pre-training with general-purpose image databases and fine-tuning with application-specific data are often used [131]. Other deep architectures that have recently been applied in remote sensing [40, 83, 123, 210] are based on deep belief networks, deep Boltzmann machines, long short-term memory networks, and stacked auto-encoders among others [80].

Kernel machines also are a well-established class of learning methods [46] that have been applied to multiple problems of remote sensing image analysis. They are the result of the combination of methodological contributions from Hilbert space theory [189], statistical learning theory [201], and quadratic optimization (see also Sect. 1.3.5).

A *kernel function* is a function of two vectors that is equivalent to evaluating an inner product in a separable Hilbert space [189]. A key idea common to all kernel machines is to make use of such kernel functions to nonlinearly generalize linear algorithms that can be entirely formulated in terms of dot

products in the Euclidean n-dimensional space—a procedure named *kernel trick*. These nonlinear extensions are methodologically equivalent to the application of the original linear techniques in a nonlinearly transformed (possibly infinite-dimensional) Hilbert space, thus gaining in flexibility and in robustness to curse-of-dimensionality [46, 201].

Analytical conditions for a function to be a kernel are well known from the theory of integral equations in functional analysis [46]. *Support vector machines* (SVM) are a well-known example of kernel machines that have been widespread and successful in remote sensing applications for over a decade [30, 46]. SVMs integrate kernels with learning criteria that optimize generalization capability and that are related to the Vapnik-Chervonenkis [201] and probably approximately correct (PAC) learning theories [46]. Numerous approaches based on the kernel and SVM frameworks have been developed in remote sensing for classification, regression, data transformation, change detection, target detection, and nonparametric density estimation [16, 30, 138, 147, 182]. Specific kernel functions and methods devoted to incorporating probabilistic graphical models, multisource information, manifolds, and sparsity were also proposed (see also Chap. 10) [79, 121, 195]. Moreover, the Gaussian process regression, which was recalled in Sect. 1.3.3 among the Bayesian nonparametric techniques, can also be interpreted as a kernel method whose kernel is the autocovariance function of the considered multidimensional Gaussian process and which formalizes a mean-square-error regression in a suitable Hilbert space [166].

Futhermore, graph-theoretic concepts, besides their relevance for data representation, are also at the basis of a further important family of methods for classification and regression with remote sensing imagery [35, 133, 145].

A *tree classifier* is determined by a finite set of decision rules that are connected and sequentially applied according to a tree topology [26]. The rule associated with each node of the tree determines a "split" of the classification procedure into two or more branches. The intuitive idea is that even though the individual splits are often very simple rules, their overall arrangement in the tree can provide flexible and powerful decision criteria.

A further topical extension is the *random forest* classifier, which has been extensively used in remote sensing. It brings tree classifiers together with the *ensemble learning* approach, in which the outputs of multiple, possibly weak, methods are combined into a unique, stronger, technique [102]. Specifically, a random forest is a finite ensemble of tree classifiers that have been parameterized in an independent and identically distributed way [25].

Formulations of the tree and random forest approaches for regression have been developed as well [25, 26].

Empirically defined tree classifiers, in which the tree structure and the splits are indicated by human prior knowledge on the considered problem, have been used in operational applications of satellite image analysis for long (e.g., for cloud screening or snow cover mapping) [98]. However, a tree classifier can also be constructed automatically on the basis of the input data. In this case, impurity measures defined in probabilistic or information-theoretic terms are typically used to determine the split in each node. Tree pruning may also be applied. Examples of well-known tree classifiers include ID3 (iterative dichotomiser 3), C4.5, and CART (classification and regression tree) [26, 102]. When multiple trees are used in a random forest to classify in an n-dimensional space, appropriate random samplings of both the data samples and the n variables are used to favor diversity among the trees—a property that positively affects the overall performance of the ensemble (see Chap. 3) [25, 102]. Other tree ensemble methods have also been proposed, including the rotation forest, extra tree, and gradient boosted tree approaches among others [18, 140]. More generally, ensemble approaches, regardless of the specific use of tree methods, represent a well-established field of machine learning [102] and are of primary importance in the framework of remote sensing data fusion (more comments on this aspect can be found in Chap. 7) [18, 173].

Finally, a further relevant class of mathematical models and methods for non-Bayesian learning from remote sensing imagery is rooted in fuzzy logic and fuzzy set theory [19]. They generalize the usual Boolean algebra and set theory to better capture uncertain behaviors. The notion of uncertainty plays a primary role in many contexts in relation, for example, to the inaccuracies in the measurements of deterministic quantities or to the qualitative descriptions of specific scenarios. As an alternative to probabilistic reasoning, fuzzy modeling has proven to be a useful mathematical tool to provide both qualitative and quantitative characterizations of uncertain scenarios.

Unlike classical set theory, in which an element either belongs or does not belong to a given set, *fuzzy set theory* allows for an element to belong to various sets with different degrees of membership (ranging in [0, 1]) [19]. A similar generalization is operated in terms of truth values by *fuzzy logic* as compared to Boolean logic. These key ideas have led to remarkable developments in pattern recognition including fuzzy classification and segmentation methods [19] as well as uncertainty models for the implementation of fuzzy reasoning rules into expert systems and knowledge representation techniques.

Many applications to image classification, segmentation, and fusion have been proposed in the remote sensing literature [7, 29, 63, 137, 155, 180]. The neural and fuzzy approaches have also been combined into *neuro-fuzzy* hybrid systems (e.g., fuzzy-ARTMAP) that incorporate fuzzy rules into neural architectures to benefit from both universal approximation and uncertainty modeling capabilities [7, 155]. From a mathematical perspective, the fuzzy approach has also been combined with

measure-theoretic concepts to introduce fuzzy measures and fuzzy integrals [165], which can be related to the plausibility and belief functions used for data fusion by Dempster-Shafer's theory of evidence (see Chap. 7).

1.3.5 The Role of Optimization Methods

In the previous sections, numerous maximization and minimization problems have been mentioned with regard to the decision rules, parameter estimations, hyperparameter optimizations, etc., that are involved in various image analysis tasks.

> The family of the *mathematical optimization* methods that are used in pattern recognition and image processing for remote sensing is vast and involves computational algorithms drawn from several areas of operations research, numerical analysis, and discrete mathematics. They include but are not restricted to linear, quadratic, convex, and non-convex programming, graph-theoretic, stochastic, and multiobjective optimization, bioinspired metaheuristics, and calculus of variations.

To name a few major examples, we recall that the learning process of an SVM with moderate to large input data sets is made possible by efficient case-specific *quadratic programming* algorithms, which were developed on purpose for SVM-related problems, and by *linear programming* methods in some special cases [46]. The same comment holds for many other kernel machines (see Chap. 10). The learning of several neural network architectures (e.g., multilayer perceptrons) can be formalized as the minimization of non-convex cost functions with respect to the network parameters through iterative algorithms such as *gradient descent*, *conjugate gradient*, *Newton*, and *quasi-Newton* methods [58]. Sparse compressive sensing representations would involve, in principle, NP-hard optimization problems based on ℓ_0 pseudo-norms, but tractable approximations have been introduced using ℓ_1 norms, linear, and convex programming [33, 158] (see Chap. 9).

The minimization of the energy functions of MRF and CRF models are challenging combinatorial problems with a huge number of variables. They are typically addressed using stochastic minimization techniques, such as *simulated annealing* [73], or optimization methods on graphs, such as *graph cuts* (based on the max-flow/min-cut theorem; see Chap. 4) [24, 82] and *belief propagation* [88]. In the case of marked point processes, energy minimization is again very challenging and is usually tackled through advanced stochastic methods involving *reversible jump Monte Carlo Markov chains* [54, 81, 152] and *birth-death* algorithms [17, 53, 75]. Hyperspectral unmixing problems (see Chap. 2) and fuzzy classification are typically formalized as constrained optimization problems in suitable multidimensional Euclidean spaces [19, 124].

Multiobjective optimization methods address problems in which multiple objective functions have to be minimized simultaneously. They allow appropriate optimality conditions (the so-called Pareto front) to be defined and the trade-off among the possibly conflicting objectives to be studied [137, 142]. Examples of applications in remote sensing include classification, regression, and unmixing techniques [119, 157].

Metaheuristics make use of ideas inspired by biological, ecological, and evolutionary concepts to address difficult non-convex problems [57]. They have been applied to many optimization tasks involved in classification, registration, and regression [18, 108, 137, 155, 181]. Well-known examples include genetic and memetic algorithms, particle swarm optimization, ant colony optimization, and artificial bee colony algorithms [57]. Convergence properties are analytically known in a relatively few cases, but these methods have been found effective in several complex minimization problems nonetheless.

Variational methods usually operate in a deterministic continuous-variable framework, formalize images as functions defined on subsets of a 2D Euclidean space, and formulate image processing problems as the minimization of suitable integral functionals (e.g., total variation functionals). Methodological concepts from the calculus of variations (e.g., the well-known Euler-Lagrange equations) often allow these methods to be expressed in terms of partial differential equations (e.g., anisotropic diffusion equations) [77], which are discretized and numerically solved on the pixel lattice. Examples of applications in remote sensing can be found in problems of restoration of degraded images, of segmentation, and of multisource data fusion and assimilation [62, 72, 174].

1.4 Structure and Organization of the Book

The remainder of the book is organized in nine chapters. Each chapter first provides an overview of the basic concepts, current literature, and main challenges related to the corresponding topic. Then, it presents at least two advanced remote sensing image analysis methodologies with examples of experimental results.

Chapters 2 through 5 focus on the main types of 2D remote sensing data and on related processing and analysis techniques. Chapter 2 is about hyperspectral image analysis with a special focus on models and optimization methods for unmixing and on binary partition trees for object-based classification. The review of the fundamentals of supervised classification that is presented here also sets the basic terminology used in the other chapters involving classification tasks. Chapter 3 is dedicated to VHR optical images and especially focuses on image data representation through graph-theoretic models for data mining and on tempo-angular anisotropic models for multiangular spectral signatures. In Chap. 4, VHR SAR imagery and InSAR data are addressed, probability density estimation for SAR data is discussed, and Markovian and patch-based models for SAR image estimation, denoising, and InSAR phase unwrapping are presented. Chapter 5 is dedicated to PolSAR and especially discusses

probability density estimation for matrix-valued PolSAR data through Mellin transforms and time-frequency decompositions of PolSAR images.

Chapters 6 and 7 are devoted to data fusion problems with multisource remote sensing imagery. In Chap. 6, the signal-level multiresolution fusion of panchromatic and hyperspectral images and the fusion of spectral, spatial, and elevation features are accomplished using guided filtering and graph-theoretic methodologies. In Chap. 7, multilevel feature extraction and multiresolution, multiscale, and multisensor image classification are addressed through tree image representations associated with mathematical morphology and hierarchical Markov random fields.

In Chaps. 8 and 9, the analysis of multitemporal remote sensing images is discussed. Chapter 8 is about change detection with multitemporal images and especially focuses on SAR change detection methodologies based on information-theoretic concepts and multiscale wavelet transforms. In Chap. 9, satellite image time series are considered, models for time series data representation and mining are described, and estimation-theoretic and compressive sensing techniques for missing data reconstruction are presented.

Finally, Chap. 10 is dedicated to kernel machines for classification and regression with remote sensing imagery. In particular, classification through multiple kernel learning, image representation based on manifold alignments, and biophysical parameter retrieval using Gaussian process regression are described.

As the chapters cover a broad range of subjects, specific notations will be introduced on a case-by-case basis. Among the few notations that are common to all chapters, we recall that $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$, and $\mathbb{C}$ will indicate as usual the semigroup of natural numbers, the group of integer numbers, and the fields of rational, real, and complex numbers, respectively. If $z \in \mathbb{C}$, then z^* will denote the conjugate of z. Boldface font will be used for vectors and matrices (with either real- or complex-valued entries). The superscripts "T" and "-1" will indicate matrix transpose and inverse, respectively, and the determinant of a square matrix $\mathbf{A}$ will be denoted $|\mathbf{A}|$ or $\det \mathbf{A}$. Given a probability space, the probability measure and the expectation operator will be written $P(\cdot)$ and $E\{\cdot\}$, respectively. However, $\mathbb{P}(\cdot)$ and $\mathbb{E}\{\cdot\}$ will also be used sometimes to avoid ambiguities with respect to chapter-specific symbols.

Acknowledgements The authors would like to thank the French Space Agency (CNES) and Airbus DS, the Italian Space Agency (ASI), the European Space Agency (ESA), the Canadian Space Agency (CSA) and MacDonald, Dettwiler and Associated Ltd. (MDA), the Japanese Space Agency (JAXA) and PASCO CORPORATION, the National Aeronautics and Space Administration (NASA) and the US Geological Survey (USGS), and the Purdue University for providing the Pléiades image in Fig. 1.4e, the COSMO-SkyMed image in Fig. 1.3c, the Sentinel-1/2/3 images in Figs. 1.3a, 1.4b, and 1.4d, the RADARSAT-2 image in Fig. 1.3b, the PALSAR-2 image in Fig. 1.3d, the Landsat-8 image in Fig. 1.4c, and the HYDICE image in Fig. 1.2, respectively. The RADARSAT-2 image in Fig. 1.3b was provided within the SOAR project number 5245. Regarding the aerial image in Fig. 1.4f, the authors would like to thank the Belgian Royal Military Academy for acquiring and providing the data and the IEEE GRSS Image Analysis and Data Fusion Technical Committee.

References

1. IEEE standard letter designations for radar-frequency bands. IEEE Std 521-2002 (Revision of IEEE Std 521-1984) pp. 1–3 (2003)
2. [grss_dfc_2015]: Online: http://www.grss-ieee.org/community/technical-committees/data-fusion
3. Aiazzi, B., Alparone, L., Baronti, S., Garzelli, A., Zoppetti, C.: Nonparametric change detection in multitemporal SAR images based on mean-shift clustering. IEEE Trans. Geosci. Remote Sens. **51**(4), 2022–2031 (2013)
4. Alcantara, E.: Remote Sensing: Techniques, Applications and Technologies. Nova Publisher (2013)
5. Alonso-Gonzalez, A., Valero, S., Chanussot, J., Lopez-Martinez, C., Salembier, P.: Processing multidimensional SAR and hyperspectral images with binary partition tree. Proc. IEEE **101**(3), 723–747 (2013)
6. Alparone, L., Aiazzi, B., Baronti, S., Garzelli, A.: Remote Sensing Image Fusion. CRC Press (2015)
7. Amici, G., Dell'Acqua, F., Gamba, P., Pulina, G.: A comparison of fuzzy and neuro-fuzzy data fusion for flooded area mapping using SAR images. Int. J. Remote Sens. **25**(20), 4425–4430 (2004)
8. Anfinsen, S., Eltoft, T.: Application of the matrix-variate Mellin transform to analysis of polarimetric radar images. IEEE Trans. Geosci. Remote Sens. **49**(6 PART 2), 2281–2295 (2011)
9. Argenti, F., Lapini, A., Bianchi, T., Alparone, L.: A tutorial on speckle reduction in synthetic aperture radar images. IEEE Geosci. Remote Sens. Mag. **1**(3), 6–35 (2013)
10. Ash, R.B.: Information Theory. Dover (1965)
11. Bachmann, C., Ainsworth, T., Fusina, R.: Improved manifold coordinate representations of large-scale hyperspectral scenes. IEEE Trans. Geosci. Remote Sens. **44**(10), 2786–2803 (2006)
12. Balanis, C.A.: Antenna Theory: Analysis and Design. Wiley (2016)
13. Ban, Y. (ed.): Multitemporal Remote Sensing. Springer (2016)
14. Barrett, E.C.: Introduction to Environmental Remote Sensing. Routledge (1999)
15. Barron, A.R.: Universal approximation bounds for superpositions of a sigmoidal function. IEEE Trans. Inf. Theory **39**(3), 930–945 (1993)
16. Bazi, Y., Melgani, F.: Toward an optimal SVM classification system for hyperspectral remote sensing images. IEEE Trans. Geosci. Remote Sens. **44**(11), 3374–3385 (2006)
17. Benedek, C., Descombes, X., Zerubia, J.: Building development monitoring in multitemporal remotely sensed image pairs with stochastic birth-death dynamics. IEEE Trans. Pattern Anal. Mach. Intell. **34**(1), 33–50 (2012)
18. Benediktsson, J.A., Ghamisi, P.: Spectral-Spatial Classification of Hyperspectral Remote Sensing Images. Artech House (2015)
19. Bezdek, J.C.: Pattern Recognition with Fuzzy Objective Function Algorithms. Springer (1981)
20. Bioucas-Dias, J., Plaza, A., Dobigeon, N., Parente, M., Du, Q., Gader, P., Chanussot, J.: Hyperspectral unmixing overview: Geometrical, statistical, and sparse regression-based approaches. IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens. **5**(2), 354–379 (2012)
21. Bishop, C.: Pattern Recognition and Machine Learning. Springer (2006)
22. Boissin, B., Ultre-Guerard, P.: The CNES Earth observation program. IEEE Geosci. Remote Sens. Mag. **3**(2), 41–50 (2015)
23. Bovolo, F., Bruzzone, L.: The time variable in data fusion: A change detection perspective. IEEE Geosci. Remote Sens. Mag. **3**(3), 8–26 (2015)
24. Boykov, Y., Veksler, O., Zabih, R.: Fast approximate energy minimization via graph cuts. IEEE Trans. Pattern Anal. Mach. Intell. **23**(11), 1222–1239 (2001)
25. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)
26. Breiman, L., Friedman, J., Stone, C.J., Olshen, R.: Classification and Regression Trees. Chapman and Hall/CRC (1984)

27. Bruce, L.M., Cheriyadat, A., Burns, M.: Wavelets: Getting perspective. IEEE Potentials **22**(2), 24–27 (2003)
28. Bruzzone, L., Bovolo, F.: A novel framework for the design of change-detection systems for very-high-resolution remote sensing images. Proc. IEEE **101**(3), 609–630 (2013)
29. Campbell, J.B., Wynne, R.H.: Introduction to Remote Sensing. Guilford Press (2011)
30. Camps-Valls, G., Bruzzone, L. (eds.): Kernel Methods for Remote Sensing Data Analysis. Wiley (2009)
31. Camps-Valls, G., Tuia, D., Gomez-Chova, L., Jimenez, S., Malo, J.: Remote Sensing Image Processing. Morgan and Claypool (2011)
32. Camps-Valls, G., Verrelst, J., Munoz-Mari, J., Laparra, V., Mateo-Jimenez, F., Gomez-Dans, J.: A survey on Gaussian processes for Earth-observation data analysis: A comprehensive investigation. IEEE Geosci. Remote Sens. Mag. **4**(2), 58–78 (2016)
33. Candes, E., Romberg, J., Tao, T.: Stable signal recovery from incomplete and inaccurate measurements. Commun. Pure Appl. Math. **59**(8), 1207–1223 (2006)
34. Canty, M.J.: Image Analysis, Classification and Change Detection in Remote Sensing: With Algorithms for ENVI/IDL and Python. Guilford Press (2014)
35. Cavallaro, G., Dalla Mura, M., Benediktsson, J., Plaza, A.: Remote sensing image classification using attribute filters defined over the tree of shapes. IEEE Trans. Geosci. Remote Sens. **54**(7), 3899–3911 (2016)
36. Celeux, G., Diebolt, J.: The SEM algorithm: A probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. Comput. Stat. Quart. **2**, 73–82 (1985)
37. Cerra, D., Datcu, M.: Expanding the algorithmic information theory frame for applications to Earth observation. Entropy **15**(1), 407–415 (2013)
38. Chanussot, J., Collet, C., Chehdi, K. (eds.): Multivariate Image Processing. Wiley (2009)
39. Chen, C.H. (ed.): Signal and Image Processing for Remote Sensing. CRC Press (2012)
40. Chen, Y., Zhao, X., Jia, X.: Spectral-spatial classification of hyperspectral data based on deep belief network. IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens. **8**(6), 2381–2392 (2015)
41. Cloude, S.: Polarisation: Applications in Remote Sensing. Oxford University Press (2014)
42. Collier, H. (ed.): Remote Sensing: Techniques and Applications. Syrawood Publishing House (2016)
43. Corsini, G., Diani, M., Grasso, R., De Martino, M., Mantero, P., Serpico, S.: Radial basis function and multilayer perceptron neural networks for sea water optically active parameter estimation in case II waters: A comparison. Int. J. Remote Sens. **24**(20), 3917–3932 (2003)
44. Craciun, P., Ortner, M., Zerubia, J.: Joint detection and tracking of moving objects using spatio-temporal marked point processes. In: IEEE Winter Conference on Applications of Computer Vision. Hawaii, USA (2015)
45. Cracknell, A.P.: Introduction to Remote Sensing. CRC Press (2007)
46. Cristianini, N., Shawe-Taylor, J.: An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. Cambridge University Press (2000)
47. Cybenko, G.: Approximation by superpositions of a sigmoidal function. Math. Control Signals. Syst. **2**(4), 303–314 (1989)
48. Datcu, M., Daschiel, H., Pelizzari, A., Quartulli, M., Galoppo, A., Colapicchioni, A., Pastori, M., Seidel, K., Marchetti, P., D'Elia, S.: Information mining in remote sensing image archives: System concepts. IEEE Trans. Geosci. Remote Sens. **41**(12 PART I), 2923–2936 (2003)
49. Datcu, M., Melgani, F., Piardi, A., Serpico, S.: Multisource data classification with dependence trees. IEEE Trans. Geosci. Remote Sens. **40**(3), 609–617 (2002)
50. de Jong, S.M., van der Meer, F.D. (eds.): Remote Sensing Image Analysis: Including the Spatial Domain. Springer (2004)
51. Deledalle, C.A., Denis, L., Poggi, G., Tupin, F., Verdoliva, L.: Exploiting patch similarity for SAR image processing: The nonlocal paradigm. IEEE Signal Process. Mag. **31**(4), 69–78 (2014)
52. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. J. Royal Stat. Soc. Series B (Methodological) **39**(1), 1–38 (1977)
53. Descombes, X. (ed.): Stochastic Geometry for Image Analysis. Wiley (2011)

54. Descombes, X., Zerubia, J.: Marked point process in image analysis. IEEE Signal Process. Mag. **19**(5), 77–84 (2002)
55. Diestel, R.: Graph Theory. Springer (2017)
56. Donoho, D.: Compressed sensing. IEEE Trans. Inf. Theory **52**(4), 1289–1306 (2006)
57. Dreo, J., Petrowski, A., Siarry, P., Taillard, E.: Metaheuristics for Hard Optimization. Springer (2006)
58. Du, K.L., Swamy, M.N.S.: Neural Networks and Statistical Learning. Springer (2014)
59. Duquenoy, M., Ovarlez, J., Ferro-Famil, L., Pottier, E., Vignaud, L.: Scatterers characterisation in radar imaging using joint time-frequency analysis and polarimetric coherent decompositions. IET Radar Sonar Nav. **4**(3), 384–402 (2010)
60. Elachi, C., van Zyl, J.J.: Introduction to the Physics and Techniques of Remote Sensing. Wiley (2006)
61. Emery, W.J., Camps, A.: Introduction to Satellite Remote Sensing. Elsevier (2017)
62. Entekhabi, D.: Land Surface Remote Sensing. SPIE Press (2013)
63. Foody, G.: The continuum of classification fuzziness in thematic mapping. Photogramm. Eng. Remote Sens. **65**(4), 443–451 (1999)
64. Foody, G.: Supervised image classification by MLP and RBF neural networks with and without an exhaustively defined set of classes. Int. J. Remote Sens. **25**(15), 3091–3104 (2004)
65. Foody, G.M., Atkinson, P.M. (eds.): Uncertainty in Remote Sensing and GIS. Wiley (2002)
66. Franceschetti, G., Lanari, R.: Synthetic Aperture Radar Processing. CRC Press (1999)
67. Frery, A.C., Müller, H.J., Yanasse, C.D.C.F., Sant'Anna, S.J.S.: A model for extremely heterogeneous clutter. IEEE Trans. Geosci. Remote Sens. **35**(3), 648–659 (1997)
68. Fujii, T., Fukuchi, T.: Laser Remote Sensing. CRC Press (2005)
69. Fukunaga, K.: Introduction to Statistical Pattern Recognition. Academic Press (2011)
70. Gabarda, S., Cristobal, G.: Cloud covering denoising through image fusion. Image Vision Comput. **25**(5), 523–530 (2007)
71. Gamba, P., Herold, M. (eds.): Global Mapping of Human Settlement: Experiences, Datasets, and Prospects. CRC Press (2009)
72. Garzelli, A.: A review of image fusion algorithms based on the super-resolution paradigm. Remote Sens. **8**(10), 1 (2016)
73. Geman, S., Geman, D.: Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Trans. Pattern Anal. Mach. Intell. **PAMI-6**(6), 721–741 (1984)
74. Gerke, M., Butenuth, M., Heipke, C., Willrich, F.: Graph-supported verification of road databases. ISPRS J. Photogramm. Remote Sens. **58**(3–4), 152–165 (2004)
75. Geyer, C.J., Møller, J.: Simulation procedures and likelihood inference for spatial point processes. Scand. J. Stat. **21**(4), 359–373 (1994)
76. Ghamisi, P., Dalla Mura, M., Benediktsson, J.: A survey on spectral-spatial classification techniques based on attribute profiles. IEEE Trans. Geosci. Remote Sens. **53**(5), 2335–2353 (2015)
77. Giaquinta, M., Hildebrandt, S.: Calculus of Variations I. Springer (2004)
78. Gibson, P., Power, C.: Introductory Remote Sensing Principles and Concepts. Routledge (2001)
79. Gomez-Chova, L., Tuia, D., Moser, G., Camps-Valls, G.: Multimodal classification of remote sensing images: A review and future directions. Proc. IEEE **103**(9), 1560–1584 (2015)
80. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press (2016)
81. Green, P.: Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika **82**(4), 711–732 (1995)
82. Greig, D.M., Porteous, B.T., Seheult, A.H.: Exact maximum a posteriori estimation for binary images. J. Royal Stat. Soc. Series B (Methodological) **51**(2), 271–279 (1989)
83. Han, J., Zhang, D., Cheng, G., Guo, L., Ren, J.: Object detection in optical remote sensing images based on weakly supervised learning and high-level feature learning. IEEE Trans. Geosci. Remote Sens. **53**(6), 3325–3337 (2015)
84. Hedhli, I., Moser, G., Zerubia, J., Serpico, S.: A new cascade model for the hierarchical joint classification of multitemporal and multiresolution remote sensing data. IEEE Trans. Geosci. Remote Sens. **54**(11), 6333–6348 (2016)

85. Hilland, J., Stuhr, F., Freeman, A., Imel, D., Shen, Y., Jordan, R., Caro, E.: Future NASA spaceborne SAR missions. IEEE Aerosp. Electron. Syst. Mag. **13**(11), 9–16 (1998)
86. Hoberg, T., Rottensteiner, F., Feitosa, R., Heipke, C.: Conditional random fields for multitemporal and multiscale classification of optical satellite imagery. IEEE Trans. Geosci. Remote Sens. **53**(2), 659–673 (2015)
87. Hodges, R.P.: Underwater Acoustics: Analysis. Wiley, Design and Performance of Sonar (2010)
88. Ihler, A., Fisher III, J., Willsky, A.: Loopy belief propagation: Convergence and effects of message errors. J. Mach. Learn. Res. **6** (2005)
89. Inglada, J., Mercier, G.: A new statistical similarity measure for change detection in multitemporal SAR images and its extension to multiscale change analysis. IEEE Trans. Geosci. Remote Sens. **45**(5), 1432–1445 (2007)
90. Jain, A.K.: Fundamentals of Digital Image Processing. Prentice Hall (1989)
91. Jasani, B., Pesaresi, M., Schneiderbauer, S., Zeug, G. (eds.): Remote Sensing from Space. Supporting International Peace and Security. Springer (2009)
92. Jian, P., Chen, K., Zhang, C.: A hypergraph-based context-sensitive representation technique for VHR remote-sensing image change detection. Int. J. Remote Sens. **37**(8), 1814–1825 (2016)
93. Jones, R.: Connected filtering and segmentation using component trees. Comput. Vis. Image Underst. **75**(3), 215–228 (1999)
94. Joseph, G.: Fundamentals of Remote Sensing. Universities Press (2005)
95. Justice, C., Townshend, J., Vermote, E., Masuoka, E., Wolfe, R., Saleous, N., Roy, D., Morisette, J.: An overview of MODIS land data processing and product status. Remote Sens. Environ. **83**(1–2), 3–15 (2002)
96. Kato, Z., Zerubia, J.: Markov random fields in image segmentation. Found. Trends Signal Process. **5**(1–2), 1–155 (2012)
97. Kendall, M.G.: The Advanced Theory of Statistics. Charles Griffin and Co. (1946)
98. Khorram, S., van der Wiele, C., Koch, F., Nelson, S., Potts, M.: Principles of Applied Remote Sensing. Springer (2016)
99. Koller, D., Friedman, N.: Probabilistic Graphical Models. MIT Press (2009)
100. Koralov, L., Sinai, Y.G.: Theory of Probability and Random Processes. Springer (2013)
101. Krylov, V., Moser, G., Serpico, S., Zerubia, J.: On the method of logarithmic cumulants for parametric probability density function estimation. IEEE Trans. Image Process. **22**(10), 3791–3806 (2013)
102. Kuncheva, L.I.: Combining Pattern Classifiers: Methods and Algorithms. Wiley (2014)
103. Kunzer, C., Dech, S. (eds.): Thermal Infrared Remote Sensing. Springer (2014)
104. Kunzer, C., Dech, S., Wagner, W. (eds.): Remote Sensing Time Series. Springer (2015)
105. Landgrebe, D.A.: Signal Theory Methods in Multispectral Remote Sensing. Wiley (2003)
106. Lavender, S., Lavender, A.: Practical Handbook of Remote Sensing. CRC Press (2015)
107. Le Cun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**(7553), 436–444 (2015)
108. Le Hegarat-Mascle, S., Kallel, A., Descombes, X.: Ant colony optimization for image regularization based on a nonstationary Markov modeling. IEEE Trans. Image Process. **16**(3), 865–878 (2007)
109. Le Moigne, J., Netanyahu, N.S., Eastman, R.D. (eds.): Image Registration for Remote Sensing. Cambridge University Press (2011)
110. Lee, J.: Introduction to Topological Manifolds. Springer (2011)
111. Lee, J.S., Pottier, E.: Polarimetric Radar Imaging: From Basics to Applications. CRC Press (2009)
112. Li, D., Wang, S., Li, D.: Spatial Data Mining. Springer (2015)
113. Li, F., Xu, L., Siva, P., Wong, A., Clausi, D.: Hyperspectral image classification with limited labeled training samples using enhanced ensemble learning and conditional random fields. IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens. **8**(6), 2427–2438 (2015)
114. Li, H.C., Celik, T., Longbotham, N., Emery, W.: Gabor feature based unsupervised change detection of multitemporal SAR images based on two-level clustering. IEEE Geosci. Remote Sens. Lett. **12**(12), 2458–2462 (2015)

115. Li, J., Bioucas-Dias, J., Plaza, A.: Semisupervised hyperspectral image segmentation using multinomial logistic regression with active learning. IEEE Trans. Geosci. Remote Sens. **48**(11), 4085–4098 (2010)
116. Li, S.: Markov Random Field Modeling in Image Analysis. Springer (2009)
117. Liang, S. (ed.): Comprehensive Remote Sensing. Elsevier (2017)
118. Lillesand, T., Kiefer, R.W., Chipman, J.: Remote Sensing and Image Interpretation. Wiley (2015)
119. Lorenzi, L., Melgani, F., Mercier, G.: Missing-area reconstruction in multispectral images under a compressive sensing perspective. IEEE Trans. Geosci. Remote Sens. **51**(7), 3998–4008 (2013)
120. Loveland, T., Irons, J.: Landsat 8: The plans, the reality, and the legacy. Remote Sens. Environ. **185**, 1–6 (2016)
121. Lunga, D., Prasad, S., Crawford, M., Ersoy, O.: Manifold-learning-based feature extraction for classification of hyperspectral data: A review of advances in manifold learning. IEEE Signal Process. Mag. **31**(1), 55–66 (2014)
122. Ly, N., Du, Q., Fowler, J.: Sparse graph-based discriminant analysis for hyperspectral imagery. IEEE Trans. Geosci. Remote Sens. **52**(7), 3872–3884 (2014)
123. Lyu, H., Lu, H., Mou, L.: Learning a transferable change rule from a recurrent neural network for land cover change detection. Remote Sens. **8**(6) (2016)
124. Ma, W.K., Bioucas-Dias, J., Chan, T.H., Gillis, N., Gader, P., Plaza, A., Ambikapathi, A., Chi, C.Y.: A signal processing perspective on hyperspectral unmixing: Insights from remote sensing. IEEE Signal Process. Mag. **31**(1), 67–81 (2014)
125. Ma, Y., Fu, Y.: Manifold Learning Theory and Applications. CRC Press (2011)
126. Mallat, S.: A theory for multiresolution signal decomposition: The wavelet representation. IEEE Trans. Pattern Anal. Mach. Intell. **11**(7), 674–693 (1989)
127. Mallat, S.: A Wavelet Tour of Signal Processing: The Sparse Way. Academic Press (2009)
128. Manakos, I., Braun, M. (eds.): Land Use and Land Cover Mapping in Europe. Springer (2014)
129. Manolakis, D.G., Lockwood, R.B., Cooley, T.W.: Hyperspectral Imaging Remote Sensing. Cambridge University Press (2016)
130. Maral, G., Bousquet, M.: Satellite Communications Systems: Systems. Techniques and Technology, Wiley (2009)
131. Marmanis, D., Datcu, M., Esch, T., Stilla, U.: Deep learning Earth observation classification using ImageNet pretrained networks. IEEE Geosci. Remote Sens. Lett. **13**(1), 105–109 (2016)
132. Marroquin, J., Mitter, S., Poggio, T.: Probabilistic solution of ill-posed problems in computational vision. J. Am. Stat. Assoc. **82**(397), 76–89 (1987)
133. Mascaro, J., Asner, G., Knapp, D., Kennedy-Bowdoin, T., Martin, R., Anderson, C., Higgins, M., Chadwick, K.: A tale of two forests: Random forest machine learning aids tropical forest carbon mapping. PLoS ONE **9**(1) (2014)
134. Masi, G., Cozzolino, D., Verdoliva, L., Scarpa, G.: Pansharpening by convolutional neural networks. Remote Sens. **8**(7) (2016)
135. Massonnet, D., Souyris, J.C.: Imaging with Synthetic Aperture Radar. EPFL Press distributed by CRC Press (2008)
136. Matteoli, S., Diani, M., Corsini, G.: A tutorial overview of anomaly detection in hyperspectral images. IEEE Aerosp. Electron. Syst. Mag. **25**(7 PART 2), 5–27 (2010)
137. Maulik, U., Bandyopadhyay, S., Mukhopadhyay, A.: Multiobjective Genetic Algorithms for Clustering. Springer (2011)
138. Maulik, U., Chakraborty, D.: Remote sensing image classification: A survey of support-vector-machine-based advanced techniques. IEEE Geosci. Remote Sens. Mag. **5**(1), 33–52 (2017)
139. Méger, N., Rigotti, C., Pothier, C.: Swap randomization of bases of sequences for mining satellite image times series. Lecture notes in computer science, Proc. of the 2015 Machine Learning and Knowledge Discovery in Databases European conference, Porto, Portugal, Part II, **9285** 190–205

140. Merentitis, A., Debes, C.: Many hands make light work - on ensemble learning techniques for data fusion in remote sensing. IEEE Geosci. Remote Sens. Mag. **3**(3), 86–99 (2015)
141. Meyer, Y.: Wavelets and Operators. Cambridge Studies in Advanced Mathematics. vol. 1 (1995)
142. Miettinen, K.: Nonlinear Multiobjective Optimization. Springer (1998)
143. Mishchenko, M.I., Travis, L.D., Lacis, A.A.: Scattering, Absorption, and Emission of Light by Small Particles. Cambridge University Press (2002)
144. Moser, G., Serpico, S., Benediktsson, J.: Land-cover mapping by Markov modeling of spatial-contextual information in very-high-resolution remote sensing images. Proc. IEEE **101**(3), 631–651 (2013)
145. Mulder, V., de Bruin, S., Schaepman, M., Mayr, T.: The use of remote sensing in soil and terrain mapping - a review. Geoderma **162**(1–2), 1–19 (2011)
146. Narayan, L.R.A.: Remote sensing and its Applications. Universities Press (2014)
147. Nasrabadi, N.: Hyperspectral target detection: An overview of current and future challenges. IEEE Signal Process. Mag. **31**(1), 34–44 (2014)
148. Navulur, K., Pacifici, F., Baugh, B.: Trends in optical commercial remote sensing industry. IEEE Geosci. Remote Sens. Mag. **1**(4), 57–64 (2013)
149. Nicolas, J.M.: Introduction aux statistiques de deuxiéme espéce: Applications des log-moments et des log-cumulants á l'analyze des lois d'images radar. Trait. Signal. **19**(11), 139–167 (2002)
150. Nowozin, S., Lampert, C.: Structured learning and prediction in computer vision. Found. Trends Comput. Graphics Vis. **6**(3–4), 185–365 (2010)
151. Oliver, C., Quegan, S.: Understanding Synthetic Aperture Radar Images. SciTech Publishing (2004)
152. Ortner, M., Descombe, X., Zerubia, J.: A marked point process of rectangles and segments for automatic analysis of digital elevation models. IEEE Trans. Pattern Anal. Mach. Intell. **30**(1), 105–119 (2008)
153. Othman, E., Bazi, Y., Alajlan, N., Alhichri, H., Melgani, F.: Using convolutional features and a sparse autoencoder for land-use scene classification. Int. J. Remote Sens. **37**(10), 2149–2167 (2016)
154. Pacifici, F., Del Frate, F., Solimini, C., Emery, W.: Neural networks for land cover applications. Studies Comput. Intell. **133**, 267–293 (2008)
155. Pal, S.K., Ghosh, A., Kundu, M.K. (eds.): Soft Computing for Image Processing. Springer (2000)
156. Pasolli, E., Melgani, F., Donelli, M.: Gaussian process approach to buried object size estimation in GPR images. IEEE Geosci. Remote Sens. Lett. **7**(1), 141–145 (2010)
157. Pasolli, L., Notarnicola, C., Bruzzone, L.: Multi-objective parameter optimization in support vector regression: General formulation and application to the retrieval of soil moisture from remote sensing data. IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens. **5**(5), 1495–1508 (2012)
158. Patel, V.M., Chellappa, R.: Sparse Representations and Compressive Sensing for Imaging and Vision. Springer (2013)
159. Pesaresi, M., Benediktsson, J.: A new approach for the morphological segmentation of high-resolution satellite imagery. IEEE Trans. Geosci. Remote Sens. **39**(2), 309–320 (2001)
160. Petrou, M., Petrou, C.: Image Processing: The Fundamentals. Wiley (2011)
161. Pieczynski, W.: Multisensor triplet Markov chains and theory of evidence. Int. J. Approx. Reason. **45**(1), 1–16 (2007)
162. Plaza, A.J., Chang, C.I. (eds.): High Performance Computing in Remote Sensing. Chapman and Hall/CRC (2007)
163. Prasad, S., Bruce, L.M., Chanussot, J. (eds.): Optical Remote Sensing. Springer (2011)
164. Prost, G.L.: Remote Sensing for Geoscientists: Image Analysis and Integration. CRC Press (2013)
165. Ralescu, D., Adams, G.: The fuzzy integral. J. Math. Anal. Appl. **75**(2), 562–570 (1980)
166. Rasmussen, C.E., Williams, C.: Gaussian Processes for Machine Learning. MIT Press (2006)

167. Rees, W.G.: Physical Principles of Remote Sensing. Cambridge University Press (2012)
168. Richards, J.A.: Remote Sensing with Imaging Radar. Springer (2009)
169. Richards, J.A.: Remote Sensing Digital Image Analysis. Springer (2013)
170. Romero, A., Gatta, C., Camps-Valls, G.: Unsupervised deep feature extraction for remote sensing image classification. IEEE Trans. Geosci. Remote Sens. **54**(3), 1349–1362 (2016)
171. Rudin, W.: Fourier Analysis on Groups. Wiley (2011)
172. Sabins, F.F.: Remote Sensing: Principles and Applications. Waveland Press (2007)
173. Samat, A., Du, P., Ali Baig, M., Chakravarty, S., Cheng, L.: Ensemble learning with multiple classifiers and polarimetric features for polarized SAR image classification. Photogramm. Eng. Remote Sens. **80**(3), 239–251 (2014)
174. Samson, C., Blanc-Féraud, L., Aubert, G., Zerubia, J.: A variational model for image classification and restoration. IEEE Trans. Pattern Anal. Mach. Intell. **22**(5), 460–472 (2000)
175. Scarpa, G., Gaetano, R., Haindl, M., Zerubia, J.: Hierarchical multiple Markov chain model for unsupervised texture segmentation. IEEE Trans. Image Process. **18**(8), 1830–1843 (2009)
176. Schanda, E.: Physical Fundamentals of Remote Sensing. Springer (1986)
177. Schindler, K.: An overview and comparison of smooth labeling methods for land-cover classification. IEEE Trans. Geosci. Remote Sens. **50**(11 PART1), 4534–4545 (2012)
178. Schmetz, J., Pili, P., Tjemkes, S., Just, D., Kerkmann, J., Rota, S., Ratier, A.: An introduction to Meteosat Second Generation (MSG). Bull. Am. Meteorol. Soc. **83**(7), 977–992 (2002)
179. Schott, J.R.: Remote Sensing: The Image Chain Approach. Oxford University Press (2007)
180. Schowengerdt, R.: Remote Sensing. Academic Press (2006)
181. Senthilnath, J., Yang, X.S., Benediktsson, J.: Automatic registration of multi-temporal remote sensing images based on nature-inspired techniques. Int. J. Image Data Fus. **5**(4), 263–284 (2014)
182. Serpico, S., Dellepiane, S., Boni, G., Moser, G., Angiati, E., Rudari, R.: Information extraction from remote sensing images for flood monitoring and damage evaluation. Proc. IEEE **100**(10), 2946–2970 (2012)
183. Serra, J.: Image Analysis and Mathematical Morphology. Academic Press (1982)
184. Sirmacek, B., Unsalan, C.: Object Detection in Satellite and Aerial Images: Remote Sensing Applications. VDM Publishing (2010)
185. Soille, P.: Morphological Image Analysis. Springer (2004)
186. Solberg, A.H.S., Taxt, T., Jain, A.K.: A Markov random field model for classification of multisource satellite imagery. IEEE Trans. Geosci. Remote Sens. **34**(1), 100–113 (1996)
187. Solimini, D.: Understanding Earth Observation. Springer (2016)
188. Stamnes, K., Thomas, G.E., Stamnes, J.J.: Radiative Transfer in the Atmosphere and Ocean. Cambridge University Press (2017)
189. Stein, E.M., Shakarchi, R.: Real Analysis: Measure Theory, Integration, and Hilbert Spaces. Princeton University Press (2005)
190. Sutton, C., McCallum, A.: An introduction to conditional random fields. Found. Trends Mach. Learn. **4**(4), 267–373 (2011)
191. Thenkabail, P.S. (ed.): Remote Sensing Handbook. CRC Press (2015)
192. Tournaire, O., Paparoditis, N.: A geometric stochastic approach based on marked point processes for road mark detection from high resolution aerial images. ISPRS J. Photogramm. Remote Sens. **64**(6), 621–631 (2009)
193. Touzi, R.: A review of speckle filtering in the context of estimation theory. IEEE Trans. Geosci. Remote Sens. **40**(11), 2392–2404 (2002)
194. Tsang, L., Kong, J.A., Ding, K.H.: Scattering of Electromagnetic Waves. Wiley, Theories and Applications (2000)
195. Tuia, D., Camps-Valls, G.: Kernel manifold alignment for domain adaptation. PLoS ONE **11**(2) (2016)
196. Tuia, D., Flamary, R., Barlaud, M.: Nonconvex regularization in remote sensing. IEEE Trans. Geosci. Remote Sens. **54**(11), 6470–6480 (2016)
197. Tuia, D., Munoz-Marí, J., Gómez-Chova, L., Malo, J.: Graph matching for adaptation in remote sensing. IEEE Trans. Geosci. Remote Sens. **51**(1), 329–341 (2013)

198. Tupin, F., Inglada, J., Nicolas, J.M. (eds.): Remote Sensing Imagery. Wiley (2014)
199. Ulaby, F.T., Long, D.G.: Microwave Radar and Radiometric Remote Sensing. Artech House (2015)
200. Van Trees, H.L., Bell, K.L., Tian, Z.: Detection Estimation and Modulation Theory. Part I: Detection, Estimation, and Filtering Theory. Wiley (2013)
201. Vapnik, V.: The Nature of Statistical Learning Theory. Springer (2000)
202. Vasile, G., Trouvé, E., Buzuloiu, V.: Intensity-driven adaptive-neighborhood technique for polarimetric and interferometric SAR parameters estimation. IEEE Trans. Geosci. Remote Sens. **44**(6), 1609–1620 (2006)
203. Velho, L., Frery, A.C., Gomes, J.: Image Processing for Computer Graphics and Vision. Springer (2009)
204. Verrelst, J., Camps-Valls, G., Muoz-Mar, J., Rivera, J., Veroustraete, F., Clevers, J., Moreno, J.: Optical remote sensing and the retrieval of terrestrial vegetation bio-geophysical properties - a review. ISPRS J. Photogramm. Remote Sens. **108**, 273–290 (2015)
205. Virelli, M., Coletta, A., Battagliere, M.L.: ASI COSMO-SkyMed: Mission overview and data exploitation. IEEE Geosci. Remote Sens. Mag. **2**(2), 64–66 (2014)
206. Volpi, M., Tuia, D.: Dense semantic labeling of subdecimeter resolution images with convolutional neural networks. IEEE Trans. Geosci. Remote Sens. **55**(2), 881–893 (2017)
207. Warner, T.A., Nellis, M.D., Foody, G.M.: The SAGE Handbook of Remote Sensing. SAGE Publishing (2009)
208. Weng, Q.: An Introduction to Contemporary Remote Sensing. McGraw Hill (2012)
209. Woodhouse, I.H.: Introduction to Microwave Remote Sensing. CRC Press (2005)
210. Zabalza, J., Ren, J., Zheng, J., Zhao, H., Qing, C., Yang, Z., Du, P., Marshall, S.: Novel segmented stacked autoencoder for effective dimensionality reduction and feature extraction in hyperspectral imaging. Neurocomputing **185**, 1–10 (2016)
211. Zhang, Y., De Backer, S., Scheunders, P.: Noise-resistant wavelet-based Bayesian fusion of multispectral and hyperspectral images. IEEE Trans. Geosci. Remote Sens. **47**(11), 3834–3843 (2009)
212. Zhu, X., Bamler, R.: Superresolving SAR tomography for multidimensional imaging of urban areas: Compressive sensing-based TomoSAR inversion. IEEE Signal Process. Mag. **31**(4), 51–58 (2014)
213. Zink, M., Bachmann, M., Brautigam, B., Fritz, T., Hajnsek, I., Wessel, B., Krieger, G.: TanDEM-X: The new global DEM takes shape. IEEE Geosci. Remote Sens. Mag. **2**(2), 8–23 (2014)

Chapter 2
Models for Hyperspectral Image Analysis: From Unmixing to Object-Based Classification

Emmanuel Maggiori, Antonio Plaza and Yuliya Tarabalka

Abstract The recent advances in hyperspectral remote sensing technology allow the simultaneous acquisition of hundreds of spectral wavelengths for each image pixel. This rich spectral information of the hyperspectral data makes it possible to discriminate different physical substances, leading to a potentially more accurate classification and thus opening the door to numerous new applications. Throughout the history of remote sensing research, numerous methods for hyperspectral image analysis have been presented. Depending on the spatial resolution of the images, specific mathematical models must be designed to effectively analyze the imagery. Some of these models operate at a sub-pixel level, trying to decompose a mixed spectral signature into its pure constituents, while others operate at a pixel or even object level, seeking to assign unique labels to every pixel or object in the scene. The spectral mixing of the measurements and the high dimensionality of the data are some of the challenging features of hyperspectral imagery. This chapter presents an overview of unmixing and classification methods, intended to address these challenges for accurate hyperspectral data analysis.

2.1 Introduction

Hyperspectral remote sensors allow the simultaneous acquisition of hundreds of spectral bands with narrow bandwidths for each image pixel. For example, the AVIRIS sensor (airborne visible/infrared imaging spectrometer) provides images with 224 contiguous bands with a bandwidth of 10 nm each, and the ROSIS sensor (reflec-

E. Maggiori (✉) · Y. Tarabalka
Université Côte d'Azur, Inria, BP 93, 2004, route des Lucioles, 06902
Sophia Antipolis Cedex, France
e-mail: emmauel.maggiori@inria.fr

Y. Tarabalka
e-mail: yuliya.tarabalka@inria.fr

A. Plaza
University of Extremadura, Badajoz, Spain
e-mail: aplaza@unex.es

G. Moser and J. Zerubia (eds.), *Mathematical Models for Remote Sensing Image Processing*, Signals and Communication Technology,
https://doi.org/10.1007/978-3-319-66330-2_2

tive optics system imagining spectrometer) provides 115 bands with a bandwidth of 4 nm each. In the spectral domain, pixels are represented as vectors for which each component is a measurement corresponding to specific wavelengths. The size of the vector is equal to the number of spectral bands that the sensor collects. For hyperspectral images, over a hundred of bands are typically available, while for conventional multispectral images up to ten bands are usually provided (see Chap. 3). This detailed spectral information increases the possibility of more accurately discriminating materials of interest [35]. The capabilities of hyperspectral sensors go beyond the identification of land cover, facilitating also the characterization of minerals [37], soils [17] and biodiversity [43]. Due to the increasing amount of data, the automatic analysis of hyperspectral images is then of paramount importance in remote sensing. One of the ultimate goals of remote sensing image analysis is to construct a thematic map associated to the image. Such a map indicates the elements present in the image, at every location, out of a set of possible classes of interest. These could go from physical substances to higher-level semantic objects, depending on the application.

While the spectral signatures collected at every pixel of a hyperspectral image are very detailed, they are usually a mixture of the signatures of the various materials found in their spatial vicinity [14]. Thus, if the spectrum is not *pure*, comparing a pixel's spectral signature with a set of reference signatures to identify the material is not an effective approach. In the earliest sensors, spatial resolution was low and the size of objects was comparable to the size of pixels. In such a low spatial resolution setting, spectral mixing compromises the key feature of the sensors: their ability to discriminate materials based on their spectral responses [13]. This drove much attention of the research community to the so-called *unmixing* of the spectral signatures, i.e., decomposing a mixed spectral signature into pure components. This can be seen as a sub-pixel analysis of the data. In hyperspectral images, the number of bands typically exceeds the amount of components in the mix, allowing to express the unmixing problem as an over-determined system of equations [47]. This chapter reviews one of the most common unmixing models, linear spectral unmixing, which assumes that the spectrum in a hyperspectral image is a linear combination of pure spectra.

As technology evolved, the spatial resolution of hyperspectral imagery increased, and objects of interest started to be composed of multiple pixels. Spectral mixing being less of a hassle, the assignment of a *unique* label to every pixel became an active research area, a process known as *classification* [13]. Since the assumption of classification is that pixels are pure, an unmixing technique is preferable if that property does not stand. The high-dimensional nature of hyperspectral imagery imposes certain challenges to perform classification, and conventional algorithms for multispectral images do not adapt well [68]. When there is a limited number of reference samples to train a classification system, as the number of dimensions increases (i.e., the number of spectral bands) the accuracy of the classification tends to drop. This is because the reliable estimation of statistical class parameters becomes more and more difficult as dimensionality increases. This phenomenon, the Hughes effect [59], is often referred to as the *curse of dimensionality*. A vast number of classification

techniques for hyperspectral imagery have been presented in the literature, which share the goal of attenuating the Hughes effect and accurately identifying the pixels' classes.

The first classification techniques were *pixelwise*, i.e., considering every pixel as an isolated entity and classifying it based solely on its spectrum. The next generation of techniques introduced the notion of a spatial arrangement of the pixels, with some interaction between spatially neighboring pixels at the time of classifying. This family of methods are known as *spectral-spatial* and tend to outperform purely pixelwise approaches [41]. The overall principle is to introduce a certain spatial regularity in the pixel label assignment, by incorporating information of the spatial neighbors. A third category is the so-called *object-based* analysis, which naturally emerged from the increase in the amount of pixels per object [15]. Object-based methods are spectral-spatial methods that seek to delineate readily usable objects to incorporate into other systems (such as geographic information systems). These techniques both segment the image into significant regions and label each of the segments. This chapter reviews pixelwise and spectral-spatial techniques, and described in detail a recent object-based model based on binary partition trees.

2.2 Unmixing

Spectral unmixing has been an alluring exploitation goal since the earliest days of hyperspectral image and signal processing [1, 14, 44, 106]. No matter the spatial resolution, the spectral signatures collected in natural environments are invariably a mixture of the signatures of the various materials found within the spatial extent of the ground instantaneous field view of the imaging instrument. For instance, the pixel vector labeled as "vegetation" in Fig. 2.1 may actually comprise a mixture of vegetation and soil, or different types of soil and vegetation canopies. In this case, several spectrally pure signatures (called *endmembers* in hyperspectral imaging terminology) are combined into the same (mixed) pixel. The availability of hyperspectral imagers with a number of spectral bands that exceeds the number of spectral mixture components [47] has allowed to cast the unmixing problem in terms of an overdetermined system of equations in which, given a set of *endmembers*, the actual unmixing to determine apparent abundance fractions can be defined in terms of a numerical inversion process [16].

A standard technique for spectral mixture analysis is *linear* spectral unmixing [55, 90, 92], which assumes that the collected spectra at the spectrometer can be expressed in the form of a linear combination of endmembers weighted by their corresponding abundances. It should be noted that the linear mixture model assumes minimal secondary reflections and/or multiple scattering effects in the data collection procedure, and hence the measured spectra can be expressed as a linear combination of the spectral signatures of materials present in the mixed pixel (see Fig. 2.2a). Although the linear model has practical advantages such as ease of implementation and flexibility in different applications [28], *nonlinear* spectral unmixing may best

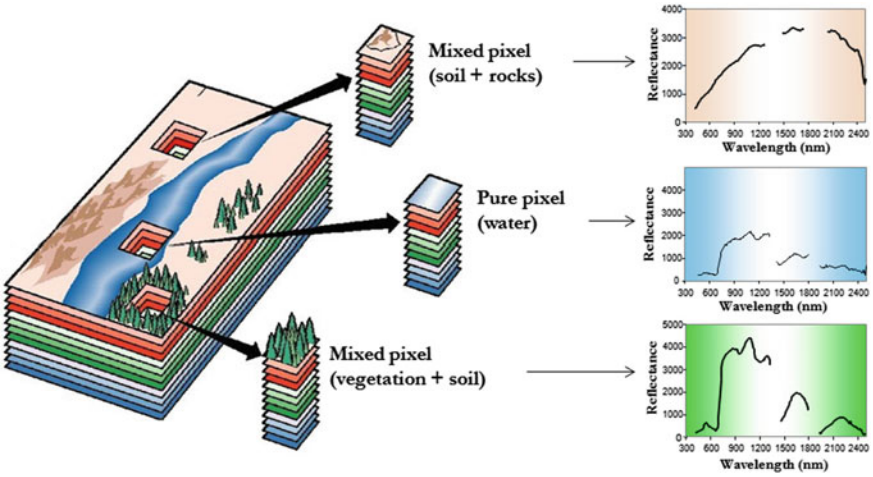


Fig. 2.1 Mixed pixels in hyperspectral imaging

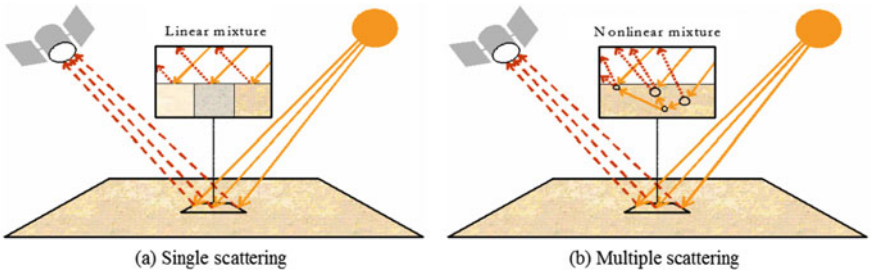


Fig. 2.2 Linear versus nonlinear mixture models: single versus multiple scattering

characterize the resultant mixed spectra for certain endmember distributions, such as those in which the endmember components are randomly distributed throughout the field of view of the instrument [49, 95]. In those cases, the mixed spectra collected at the imaging instrument is better described by assuming that part of the source radiation is multiply scattered before being collected at the sensor (see Fig. 2.2b). In this case, interactions can be at a *classical*, or *multilayered*, level or at a *microscopic*, or *intimate*, level. Mixing at the classical level occurs when light is scattered from one or more objects, is reflected off additional objects, and eventually is measured by hyperspectral imager. A nice illustrative derivation of a multilayer model is given by Borel and Gerstl [18] who show that the model results in an infinite sequence of powers of products of reflectances. Generally, however, the first order terms are sufficient and this leads to the bilinear model. Microscopic mixing occurs when two materials are homogeneously mixed [52]. In this case, the interactions consist of photons emitted from molecules of one material and absorbed by molecules of another material, which may in turn emit more photons. The mixing is modeled by Hapke [52] as occurring at the *albedo* level, i.e., the fraction of solar energy

reflected from the Earth, and not at the reflectance level (for more details on the physical quantities acquired by passive cameras see Chap. 3). The apparent albedo of the mixture is a linear average of the albedos of the individual substances but the reflectance is a nonlinear function of albedo, thus leading to a different type of nonlinear model.

In the following, we focus on describing recent advances in the linear spectral unmixing domain. The reason is that, despite its simplicity, it is an acceptable approximation of the light scattering mechanisms in many real scenarios. Furthermore, in contrast to nonlinear mixing, the linear mixing model is the basis of a plethora of unmixing models and algorithms spanning back at least 25 years. A sampling can be found in [12, 27, 54, 56, 58, 63, 64, 78, 85, 87, 90, 92, 105, 107, 130], see also [14] and references therein. As shown in Fig. 2.2a, the linear mixture model assumes that mixed pixels are a linear combination of the endmembers. This scenario holds when the mixing scale is macroscopic [108] and the incident light interacts with just one material, as is the case in checkerboard type scenes [36, 51]. In this case, the mixing occurs within the instrument itself. It is due to the fact that the resolution of the instrument is not fine enough. The light from the materials, although almost completely separated, is mixed within the measuring instrument.

In order to define the linear mixture model in mathematical terms, let us assume that $\mathbf{Y} \in \mathbb{R}^{l \times n}$ is a hyperspectral image with l bands and n pixels. In this case the matrix $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_n]$ represents a hyperspectral image in a matrix form, in which the columns of the matrix $\mathbf{Y}$ are the spectral signatures of the image pixels $\mathbf{y}_i$, and the rows of $\mathbf{Y}$ are the bands of the hyperspectral image. Under the linear mixture assumption, we can model the hyperspectral data as follows:

$$\mathbf{Y} = \mathbf{MA} + \mathbf{N}, \tag{2.1}$$

where $\mathbf{M} \in \mathbb{R}^{l \times p}$, $\mathbf{M} = [\mathbf{m}_1, \ldots, \mathbf{m}_p]$ is a matrix containing endmembers $\mathbf{m}_i$ in columns and $\mathbf{A} \in \mathbb{R}^{p \times n}$, $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_n]$ contains the abundance fractions $a_{j,k}$ associated to each endmember in each pixel. Finally, $\mathbf{N} \in \mathbb{R}^{l \times n}$ is a matrix which represents the noise introduced in the model by the acquisition process. Usually two constraints are imposed to the abundance fractions in the linear mixture model. The first one is the abundance non-negativity (ANC), which enforces to all the abundances fractions to be non-negative [31], i.e. $a_{j,k} \geq 0$, $j = 1, \ldots, p$, $k = 1, \ldots, n$. The second constraint is the abundance sum-to-one (ASC), which enforces the abundances of a given pixel to sum to one, i.e. $\sum_{j=1}^{p} a_{j,k} = 1$, $k = 1, \ldots, n$. The unmixing process which considers both constraints is called fully constrained linear spectral unmixing (FCLSU). The linear mixture model can be interpreted graphically by using a scatter plot between two bands or, more generally, between two non-colinear projections of the spectral vectors. For illustrative purposes, Fig. 2.3 provides a simple graphical interpretation in which the endmembers are the most extreme pixels defining a simplex which encloses all the other pixels in the data, so that we can express every pixel inside the simplex as a linear combination of the endmembers. As a result, a key aspect when considering the linear mixture model is the correct identification of the endmembers, which are extreme points in the l-dimensional space.

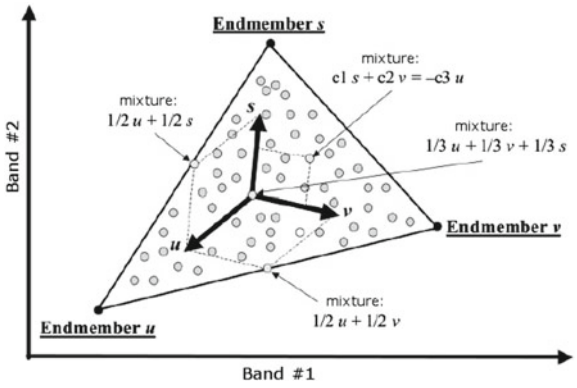


Fig. 2.3 Graphical interpretation of the linear mixture model

The solution of the linear spectral mixture problem described in (2.1) relies on two major requirements:

1. A successful estimation of how many endmembers, p, are present in the input hyperspectral scene $\mathbf{Y}$, and
2. the correct determination of a set $\mathbf{M}$ of p endmembers and their correspondent abundance fractions at each pixel.

In order to address these issues, a standard spectral unmixing chain consisting of three steps is generally applied. In a first step, an (optional) dimensionality reduction step is conducted. This step is strongly related to the estimation of the number of endmembers present in the hyperspectral scene, p. Once the number of endmembers has been determined, an endmember extraction step identifies the pure spectral signatures present in a scene. Finally, the abundance estimation step requires as input the endmember signatures obtained in the endmember extraction process and produces as output the set of abundance maps associated to each endmember. Figure 2.4 shows the different steps involved in the processing chain, which are briefly summarized next and described in more detail in the following subsections (discussing specific implementation options for each step).

1. **Dimensionality reduction**. The dimensionality of the space spanned by spectra from an image is generally much lower than the available number of bands. Identifying appropriate subspaces facilitates dimensionality reduction, improving algorithm performance and data storage complexity. Furthermore, if the linear mixture model is accurate, the signal subspace dimension is one less than the number of endmembers, a crucial figure in hyperspectral unmixing.
2. **Endmember extraction**. This step consists in identifying the endmembers in the scene. *Geometrical* approaches exploit the fact that linearly mixed vectors are in a simplex set or in a positive cone. *Statistical* approaches focus on using parameter estimation techniques to determine endmembers. Different techniques may or may not include spatial information and assume or not the presence of pure pixels in the original data set.

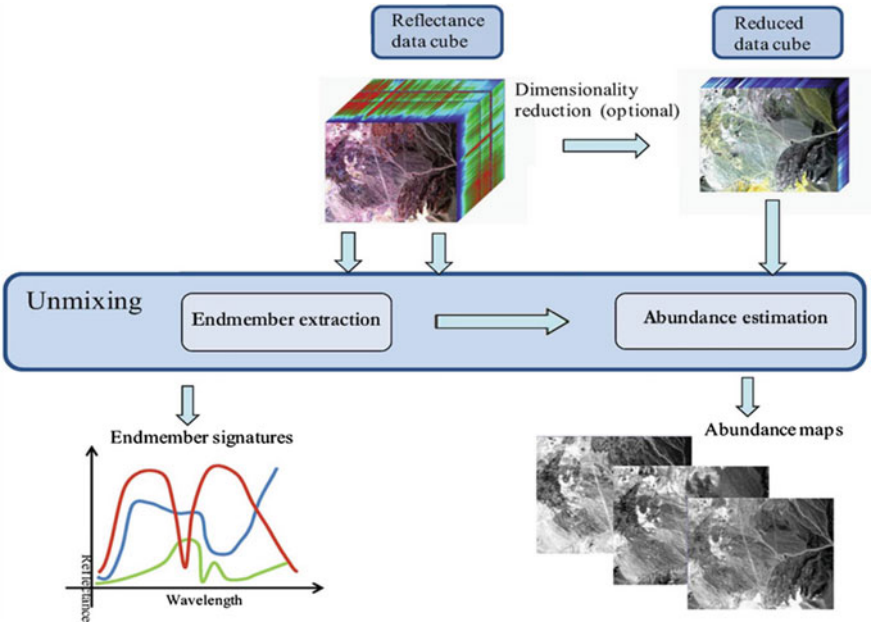


Fig. 2.4 Spectral unmixing chain

3. **Abundance estimation**. Given the identified endmembers, the abundance estimation step consists in solving a constrained optimization problem which minimizes the residual between the observed spectral vectors and the linear space spanned by the inferred endmembers in order to derive fractional abundances which are, very often, constrained to be nonnegative and to sum to one (i.e., they belong to the probability simplex). There are, however, some hyperspectral unmixing approaches in which the endmember determination and inversion steps are implemented simultaneously.

2.2.1 Dimensionality Reduction

The number of endmembers p present in a given scene is, very often, much smaller than the number of bands l. Therefore, assuming that the linear model is a good approximation, spectral vectors lie in or very close to a low-dimensional linear subspace. The identification of this subspace enables low-dimensional yet accurate representation of spectral vectors. It is usually advantageous and sometimes necessary to operate on data represented in the signal subspace. Therefore, a signal subspace identification algorithm is often required as a first processing step in the spectral unmixing chain. Unsupervised subspace identification has been addressed in many ways. Projection techniques seek for the best subspaces to represent data by optimizing

objective functions. For example, principal component analysis (PCA) maximizes the signal variance; singular value decomposition (SVD) maximizes power; minimum noise fraction (MNF) and noise-adjusted principal components (NAPC) minimize the ratio of noise power to signal power. NAPC is mathematically equivalent to MNF [70] and can be interpreted as a sequence of two principal component transforms: the first applied to the noise and the second applied to the transformed data set.

The identification of the signal subspace is a model order inference problem to which information theoretic criteria come to mind. These criteria have in fact been used in hyperspectral applications [30] adopting the approach introduced by Wax and Kailath [127]. In turn, Harsanyi, Farrand, and Chang [53] developed a Neyman-Pearson detection theory-based thresholding method to determine the number of spectral endmembers in hyperspectral data, referred to as virtual dimensionality (VD). This method is based on a detector built on the eigenvalues of the sample correlation and covariance matrices. A modified version includes a noise-whitening step [30]. The hyperspectral signal identification with minimum error (HYSIME) adopts a minimum mean squared error based approach to infer the signal subspace. The method is eigendecomposition based, unsupervised, and fully-automatic (i.e., it does not depend on any tuning parameters). It first estimates the signal and noise correlation matrices and then selects the subset of eigenvalues that best represents the signal subspace in the least square error sense.

2.2.2 Endmember Extraction

Over the last decade, several algorithms have been developed for automatic or semi-automatic extraction of spectral endmembers by assuming the presence of pure pixels in the hyperspectral data [92]. Classic techniques include the pixel purity index (PPI), N-FINDR, iterative error analysis (IEA), convex cone analysis (CCA), vertex component analysis (VCA), and orthogonal subspace projection (OSP), among many others [14]. Other advanced techniques for endmember extraction have been recently proposed [9, 26, 32, 33, 81, 89, 126, 132], but none of them considers spatial adjacency. However, one of the distinguishing properties of hyperspectral data is the multivariate information coupled with a two-dimensional (pictorial) representation amenable to image interpretation. Subsequently, most endmember extraction algorithms listed above could benefit from an integrated framework in which both the spectral information and the spatial arrangement of pixel vectors are taken into account. An example of this situation is given in Fig. 2.5, in which a hyperspectral data cube collected over an urban area (high spatial correlation) is modified by randomly permuting the spatial coordinates of the pixel vectors, thus removing the spatial correlation. In both scenes, the application of a spectral-based endmember extraction method would yield the same analysis results while it is clear that a spatial-spectral technique could incorporate the spatial information present in the original scene into the endmember searching process.

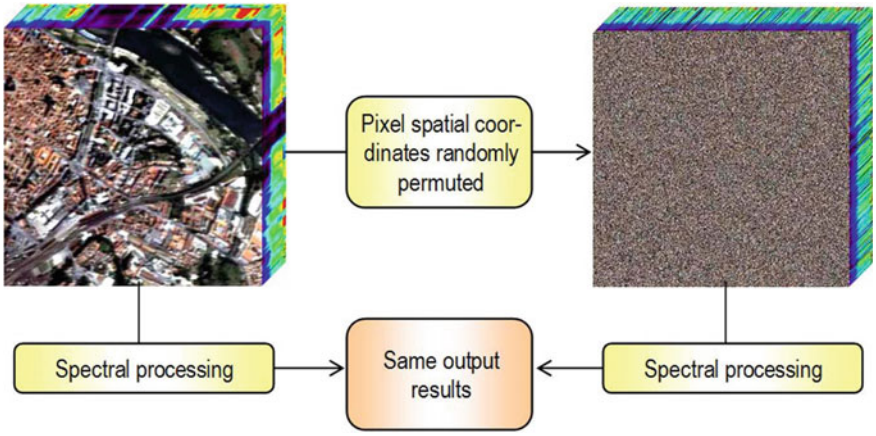
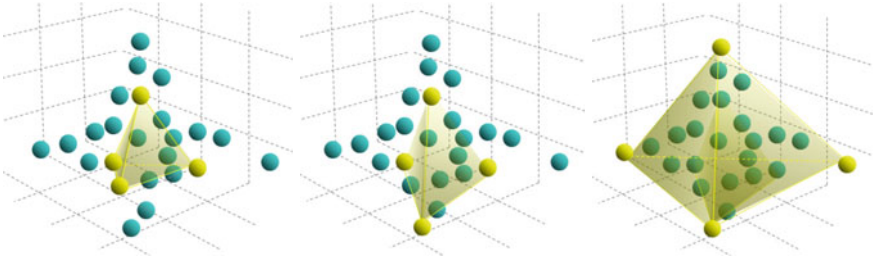


Fig. 2.5 Example illustrating the importance of spatial information in hyperspectral analysis

To the best of our knowledge, only a few attempts exist in the literature aimed at including the spatial information in the process of extracting spectral endmembers. Extended morphological operations [93] have been used as a baseline to develop the automatic morphological endmember extraction algorithm (AMEE) for spatial-spectral endmember extraction. Also, spatial averaging of spectrally similar endmember candidates found via SVD was used in the development of the spatial-spectral endmember extraction algorithm (SSEE). In the following, we describe in more detail three selected spectral-based algorithms (N-FINDR, OSP and VCA) and two spatial-spectral endmember extraction algorithms (AMEE and SSEE) that will be used in our comparisons in this chapter. The reasons for our selection are: (1) these algorithms are representative of the class of convex geometry-based and spatial processing-based techniques which have been successful in endmember extraction; (2) they are fully automated; (3) they always produce the same final results for the same input parameters; and (4) the number of endmembers to be extracted, p, is an input parameter for all algorithms, while AMEE and SSEE have additional input parameters related to the definition of spatial context around each pixel in the scene. This section concludes with a description of algorithms that, as opposed to the previously mentioned ones, do not assume the presence of pure pixels in the hyperspectral data. Techniques in this category comprise minimum volume simplex analysis (MVSA) and a variable splitting augmented Lagrangian approach (SISAL). Also, we deliberately do not cover sparse unmixing methods [60], which are detailed in other chapters of this book.

2.2.2.1 N-FINDR

This algorithm looks for the set of pixels with the largest possible volume by *inflating* a simplex inside the data. The procedure begins with a random initial selection of pixels (see Fig. 2.6a). Every pixel in the image must be evaluated in order to refine



(a) Initialization (b) Replacement (c) Final result

Fig. 2.6 Graphical representation of the N-FINDR algorithm

the estimate of endmembers, looking for the set of pixels that maximizes the volume of the simplex defined by selected endmembers. The volume of the simplex is calculated with every pixel in the place of each endmember. The corresponding volume is calculated for every pixel in each endmember position by replacing that endmember and finding the resulting volume (see Fig. 2.6b). If the replacement results in a an increase of volume, the pixel replaces the endmember. This procedure is repeated until there are no more endmember replacements (see Fig. 2.6c). The mathematical definition of the volume of a simplex formed by a set of endmember candidates is proportional to the determinant of the set augmented by a row of ones. The determinant is only defined in the case where the number of features is $p - 1$, p being the number of desired endmembers [29]. Since in hyperspectral data typically $l \gg p$, a transformation that reduces the dimensionality of the input data is required. Often, the PCA transform has been used for this purpose, although another widely used alternative that decorrelates the noise in the data is MNF. A possible shortcoming of this algorithm is that different random initializations of N-FINDR may produce different final solutions. In this chapter, we consider an N-FINDR algorithm implemented in an iterative fashion, so that each sequential run is initialized with the previous algorithm solution, until the algorithm converges to a simplex volume that cannot be further maximized.

2.2.2.2 Orthogonal Subspace Projection (OSP)

This algorithm starts by selecting the pixel vector with maximum length in the scene as the first endmember. Then, it looks for the pixel vector with the maximum absolute projection in the space orthogonal to the space linearly spanned by the initial pixel, and labels that pixel as the second endmember. A third endmember is found by applying an orthogonal subspace projection to the original image [54]. This is done by selecting the signature that has the maximum orthogonal projection in the space orthogonal to the space linearly spanned by the first two endmembers. This procedure is repeated until the desired number of endmembers, p, is found [98]. A shortcoming

of this algorithm is its sensitivity to noise, since outliers are good candidates to be selected in the iterative process adopted by OSP. The VCA method discussed in the following subsection addresses this issue.

2.2.2.3 Vertex Component Analysis (VCA)

This algorithm also makes use of the concept of orthogonal subspace projections. However, as opposed to the OSP algorithm described above, VCA exploits the fact that the endmembers are the vertices of a simplex, and that the affine transformation of a simplex is also a simplex [84]. As a result, VCA models the data using a positive cone, whose projection onto a properly chosen hyperplane is another simplex whose vertices are the final endmembers. After projecting the data onto the selected hyperplane, the VCA projects all image pixels to a random direction and uses the pixel with the largest projection as the first endmember. The other endmembers are identified in sequence by iteratively projecting the data onto a direction orthogonal to the subspace spanned by the endmembers already determined, using a procedure that is quite similar to that used by OSP. The new endmember is then selected as the pixel corresponding to the extreme projection, and the procedure is repeated until a set of p endmembers is found [84]. For illustrative purposes, Fig. 2.7 shows a toy example depicting an image with three bands and three endmembers. Due to the mixing phenomenon, all the data is in the plane $\mathbf{S}$. If we project the data onto that plane we can represent the same data in two dimensions instead of three. Then we can apply OSP to the projected dataset in order to obtain the endmembers.

A possible shortcoming of the VCA algorithm can be illustrated by the following example: if there are two endmembers with similar spectral signatures and the power of noise is high, then the subspace identification step could miss one of the two similar endmembers. This problem could be avoided by using spatial information as follows. The idea is that, although the endmembers are very similar in the spectral domain, they may be located in different areas in the spatial domain. As a result,

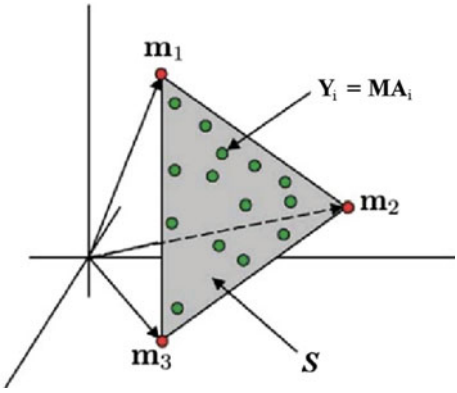


Fig. 2.7 Toy example illustrating the impact of subspace projection on endmember identification

spatial information could help in the distinction of the endmembers. In the following subsections we describe different algorithms which make use of spatial information in order to solve some of these potential problems in the endmember identification process.

2.2.2.4 Automatic Morphological Endmember Extraction (AMEE)

The AMEE [91] algorithm runs on the full data cube with no dimensional reduction, and begins by searching spatial neighborhoods around each pixel vector in the image for the most spectrally pure and mostly highly mixed pixel. This task is performed by using extended mathematical morphology operators [93] of dilation and erosion, which are graphically illustrated in Fig. 2.8. Here, dilation selects the most spectrally pure pixel in a local neighborhood around each pixel vector, while erosion selects the most highly mixed pixel in the same neighborhood. Each spectrally pure pixel is then assigned an *eccentricity* value, which is calculated as the spectral angle (SA) [28, 64] between the most spectrally pure and mostly highly mixed pixel for each given spatial neighborhood. This process is repeated iteratively for larger spatial neighborhoods up to a maximum size that is predetermined. At each iteration the eccentricity values of the selected pixels are updated. The final endmember set is obtained by applying a threshold to the resulting greyscale eccentricity image, which results in a large set of endmember candidates. The final endmembers are extracted after applying the OSP method to the set of candidates in order to derive a final set of spectrally distinct endmembers **M**, where p is an input parameter to the OSP algorithm.

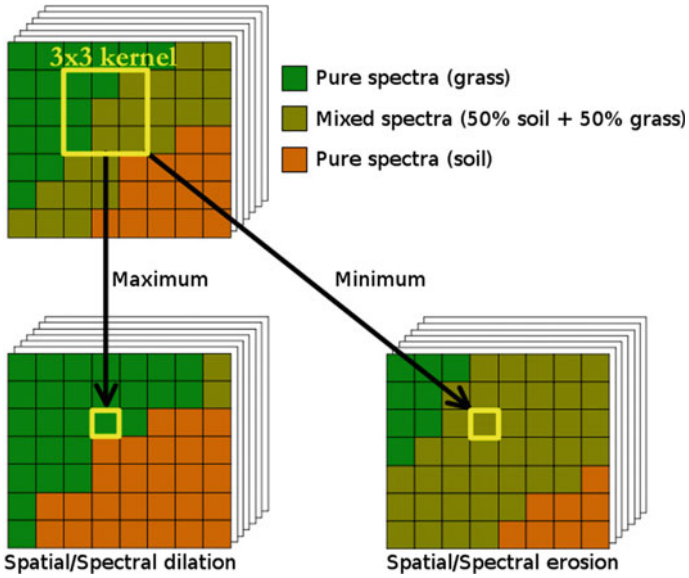


Fig. 2.8 Extended morphological operations of erosion and dilation

2.2.2.5 Spatial Spectral Endmember Extraction (SSEE)

The SSEE algorithm uses spatial constraints to improve the relative spectral contrast of endmember spectra that have minimal unique spectral information, thus improving the potential for these subtle yet potentially important endmembers to be selected. With SSEE, the spatial characteristics of image pixels are used to increase the relative spectral contrast between spectrally similar but spatially independent endmembers. The SSEE algorithm searches an image with a local search window centered around each pixel vector and comprises four steps [100]. First, the SVD transform is applied to determine a set of eigenvectors that describe most of the spectral variance in the window or partition (see Fig. 2.9). Second, the entire image data are projected onto the previously extracted eigenvectors to determine a set of candidate endmember pixels (see Fig. 2.10). Then, spatial constraints are used to combine and average spectrally similar candidate endmember pixels by testing, for each candidate pixel vector, which other pixel vectors are sufficiently similar in spectral sense (see Fig. 2.11). Instead of using a manual procedure as recommended by the authors in [100], we have used the OSP technique in order to derive a final set of spectrally distinct endmembers **M**, where p is an input parameter to the OSP algorithm.

At this point, it is important to note that SSEE includes spatial information in a different way as AMEE does. The SSEE method uses first a spectral SVD method to extract some candidate endmembers and then includes the spatial information. On the other hand, AMEE combines the spatial and spectral information at the same time using extended morphological operations, and then uses a spectral endmember extraction technique in order to select the final endmember set. In both cases (as it is also the case of all endmember identification algorithms discussed thus far) the assumption is that pure spectral signatures are present in the original hyperspectral data. In the following subsection we describe methods which operate under the assumption that pure spectral signatures may not be present at all in the original hyperspectral scene.

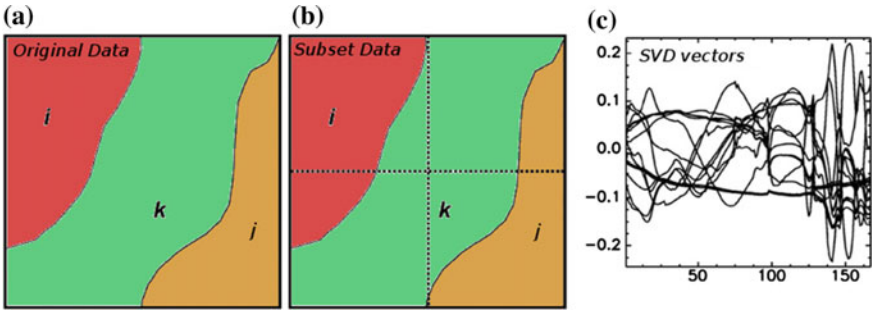


Fig. 2.9 First step of the SSEE algorithm. **A** Original data. **B** Subset data after spatial partitioning. **C** Set of representative SVD vectors used to describe spectral variance. (Figure reproduced from [100])

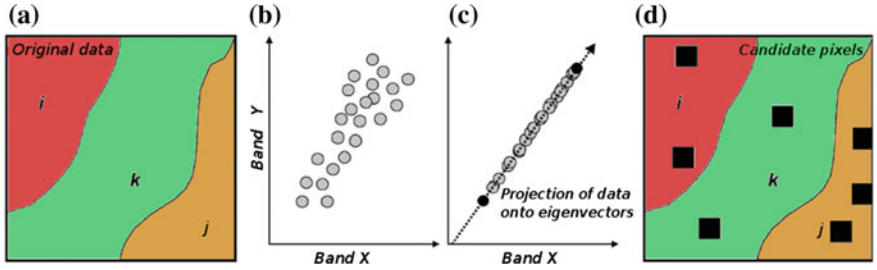


Fig. 2.10 Second step of the SSEE algorithm. **A** Original data. **B** Spectral distribution in 2-dimensional space. **C** Projection of data onto eigenvectors. **D** Set of candidate pixels. (Figure reproduced from [100])

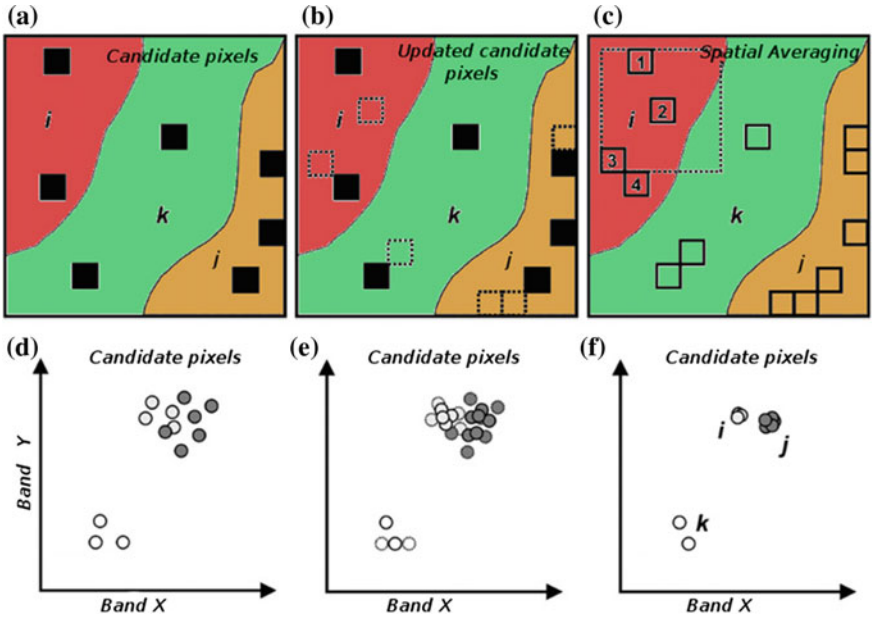
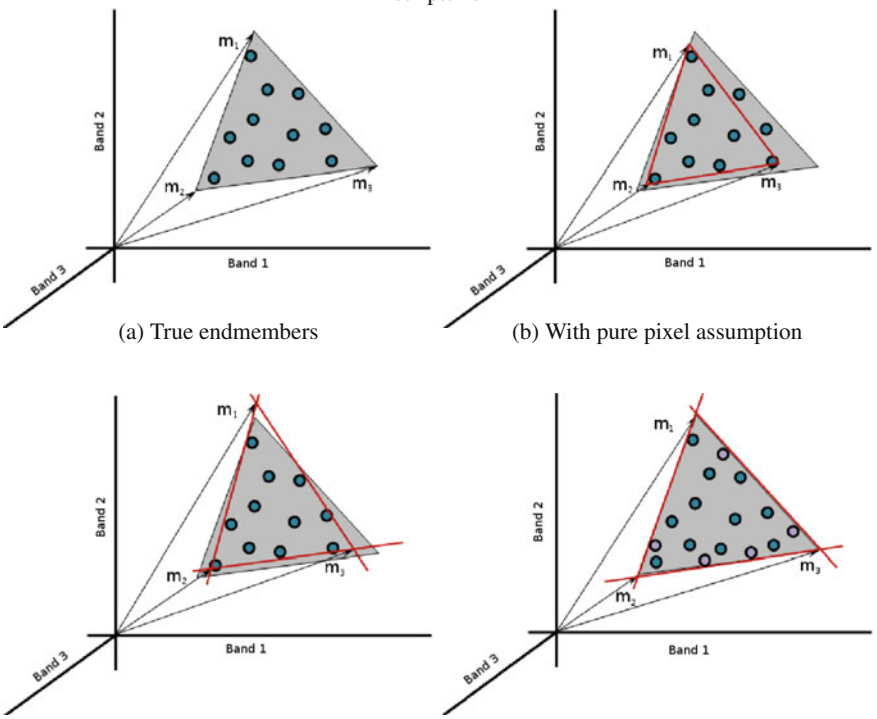


Fig. 2.11 Third step of the SSEE algorithm. **A** Set of candidate pixels. **B** Updated candidate pixels after including pixels which are spectrally similar to those in the original set. **C** Spatial averaging process of candidate endmember pixels using a sliding window centered on each candidate. **D** First iteration of spatial-spectral averaging. Averaged pixels shown as thick lines, with original pixels shown as *thinner lines*. **E** Second iteration of spatial-spectral averaging. **F** Continued iterations compress endmembers into clusters with negligible variance. (Figure reproduced from [100])

2.2.2.6 Algorithms Without the Pure Pixel Assumption

This section describes endmember identification techniques which do not operate under the pure pixel assumption [57, 94]. In this case, the algorithms do not need the presence of pure pixels in the dataset in order to generate the endmembers. Figure 2.12

scriptsize



(a) True endmembers

(b) With pure pixel assumption

(c) Without pure pixel assumption

(d) With pixels at the simplex facets

Fig. 2.12 Illustration of different strategies for endmember extraction

shows a graphical interpretation of the difference between algorithms that assume and do not assume the presence of pure pixels in the dataset. Specifically, Fig. 2.12a represents the true endmembers. In this case, there is no pixel at the simplex vertices so the endmembers are not present in the original data. Figure 2.12b represents a possible solution of an algorithm which does not assume the presence of pure pixels in the dataset. As we can see in Fig. 2.12b, there are two pixels outside of the simplex, which are outliers in this particular case. Figure 2.12c represents a possible solution of a method which does not assume the presence of pure pixels in the data. In this case, the algorithm tries to estimate a set of endmembers by enclosing the whole dataset. This approach does not guarantee the correct identification of endmembers in the case that the data are highly mixed and there are no pixels in the facets of the simplex. However, if there are pixels in the simplex facets, the true endmembers can be correctly identified even if there are no pixels at the simplex vertex, as depicted in Fig. 2.12d.

Most of the techniques in this category adopt a minimum volume strategy aimed at finding the endmember matrix **M** by minimizing the volume of the simplex defined by its columns and containing the endmembers. This is a non-convex optimization

problem much harder than those considered in the previous subsection in which the endmembers are assumed to belong to the input hyperspectral image.

Craig's seminal work [38] established the concepts regarding the algorithms of minimum volume type. Most of these algorithms formulate the endmember estimation as the nonnegative matrix factorization of the mixing and abundance matrices [9, 73, 81, 96, 131, 133], with a minimum volume constraint imposed on $\mathbf{M}$. Nonnegative matrix factorization is a hard non-convex optimization problem prone to get stuck in local minima. Aiming at obtaining lighter algorithms with more desirable convergence properties, the works [2, 10, 25, 71] sidestep the matrix factorization by formulating the endmember estimation as an optimization problem with respect to $\mathbf{Q} = \mathbf{M}^{-1}$. The MVSA and SISAL algorithms implement a robust version of the minimum volume concept. Robustness is introduced by allowing the ANC to be violated. These violations are weighted using a soft constraint given by the hinge loss function, hinge ($\mathbf{x}$), an elementwise operator that returns 0 if $x_i \geq 0$ and $-x_i$ if $x_i < 0$, for every element x_i in $\mathbf{x}$. After reducing the dimensionality of the input data from l to $p-1$, MVSA/SISAL aim at solving the following optimization problem:

$$\begin{aligned} &\widehat{\mathbf{Q}} = \arg\max_{\mathbf{Q}} \log\left(|\det\left(\mathbf{Q}\right)|\right) - \lambda \mathbf{1}_p^T \text{hinge}\left(\mathbf{QY}\right)\mathbf{1}_n \\ &\text{s.t.:} \quad \mathbf{1}_p^T \mathbf{Q} = \mathbf{q}_m, \end{aligned} \tag{2.2}$$

where $\mathbf{Q} \equiv \mathbf{M}^{-1}$, $\mathbf{1}_p$ and $\mathbf{1}_n$ are column vectors of ones of sizes p and n, respectively, $\mathbf{q}_m \equiv \mathbf{1}_p^T \mathbf{Y}_p^{-1}$ with $\mathbf{Y}_p$ being any set of linearly independent spectral vectors taken from the hyperspectral data set $\mathbf{Y}$, and λ is a regularization parameter. Here, maximizing $\log(|\det(\mathbf{Q})|)$ is equivalent to minimizing the volume of $\mathbf{M}$.

2.2.3 Abundance Estimation

Once a set of endmembers $\mathbf{M}$ have been extracted, their correspondent abundance fractions $\mathbf{A}$ can be estimated (in least squares sense) by the following unconstrained expression [28]:

$$\mathbf{A} \approx (\mathbf{M}^T\mathbf{M})^{-1}\mathbf{M}^T\mathbf{Y}. \tag{2.3}$$

However, it should be noted that the fractional abundance estimations obtained by means of Eq. (2.3) do not satisfy the ASC and ANC constraints. As indicated in [30], a non-negative constrained least squares (NCLS) algorithm can be used to obtain a solution to the ANC-constrained problem in an iterative fashion [31]. In order to take care of the ASC constraint, we replace the hard constraint $\mathbf{1}^T\mathbf{A} = 1$ by the soft constraint $\sqrt{\delta}||\mathbf{1}^T\mathbf{A} - 1||_2^2$ added to the quadratic data term $||\mathbf{Y} - \mathbf{MA}||_2^2$. This is equivalent to using a new endmember signature matrix, denoted by $\mathbf{M}'$, and a modified version of the abundance estimates $\mathbf{A}$, denoted by $\mathbf{A}'$, are introduced as follows:

$$\mathbf{M}' = \begin{bmatrix} \mathbf{M} \\ \delta \mathbf{1}^T \end{bmatrix}, \mathbf{A}' = \begin{bmatrix} \mathbf{A} \\ \delta \mathbf{1} \end{bmatrix}, \tag{2.4}$$

where $\mathbf{1} = (\underbrace{1, 1, \cdots, 1}_{p})^T$ and δ controls the impact of the ASC constraint. Using the two expressions in (2.4), a fully constrained estimate can be directly obtained from the NCLS algorithm by replacing $\mathbf{M}$ and $\mathbf{A}$ with $\mathbf{M}'$ and $\mathbf{A}'$. The fully constrained (i.e., ASC-constrained and ANC-constrained) linear spectral unmixing model is referred to as FCLSU.

2.2.4 Experimental Validation

In this section we will describe the experiments performed with a real hyperspectral dataset collected by the airborne visible infrared imaging spectrometer (AVIRIS) over the Cuprite mining district. The scene, available online in reflectance units after atmospheric correction,[1] is characterized by the availability of some very reliable reference information available from the United States Geological Survey (USGS). Specifically, the portion used in experiments corresponds to a 350 × 350-pixel subset of the sector labeled as "f970619t01p02_r02_sc03.a.rfl" in the online data. The scene comprises 224 spectral bands between 0.4 and 2.5 μm, with full width at half maximum of 10 nm and spatial resolution of 20 m per pixel. Prior to the analysis, several bands were removed due to water absorption and low signal-to-noise ratio (SNR) in those bands, leaving a total of 188 reflectance channels to be used in the experiments. The Cuprite site is well understood mineralogically, and has several exposed minerals of interest, all included in the USGS library considered in experiments, denoted "splib06" and released in September 2007.[2] In our experiments, we use spectra obtained from this library to validate endmember extraction algorithms. For illustrative purposes, Fig. 2.13 shows a mineral map produced in 1995 by USGS, in which the Tricorder 3.3 software product was used to map different minerals present in the Cuprite mining district.[3] It should be noted that the Tricorder map is only available for hyperspectral data collected in 1995, while the publicly available AVIRIS Cuprite data was collected in 1997. Therefore, a direct comparison between the 1995 USGS map and the 1997 AVIRIS data (as well as a comparison in terms of fractional abundances) is not possible.

We show a comparison of the results obtained for the endmember extraction algorithms in terms of accuracy and also in terms of computational complexity. Accuracy is measured in terms of the spectral angle (SA), i.e., the angle between two spectral signature vectors. Table 2.1 tabulates the SA scores, in degrees, obtained after comparing the USGS library spectra of *alunite*, *buddingtonite*, *calcite*, *kaolinite*

[1] http://aviris.jpl.nasa.gov/html/aviris.freedata.html.

[2] http://speclab.cr.usgs.gov/spectral.lib06.

[3] http://speclab.cr.usgs.gov/cuprite95.tgif.2.2um_map.gif.

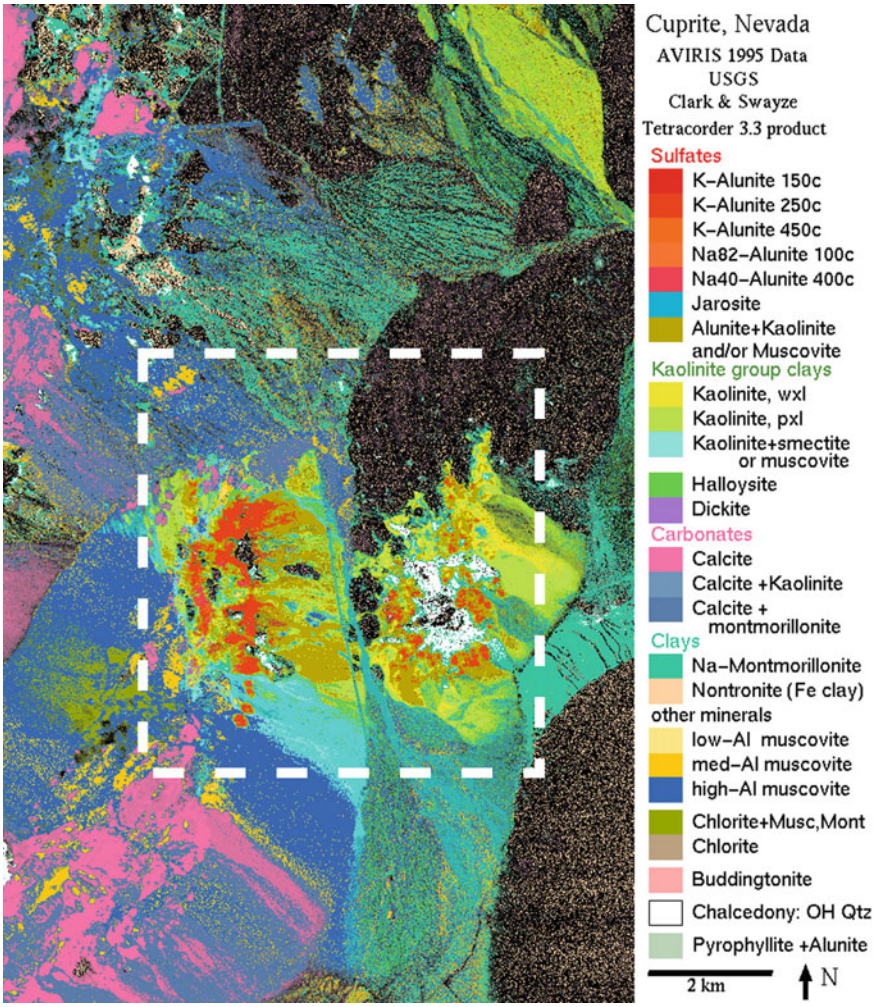


Fig. 2.13 USGS map showing the location of different minerals in the Cuprite mining district in Nevada. The map is available online at: http://speclab.cr.usgs.gov/cuprite95.tgif.2.2um_map.gif. The white rectangle depicts the area used in our experiments

and *muscovite*, with the corresponding endmembers extracted by different algorithms from the AVIRIS Cuprite scene. In all cases, the input parameters of the different endmember extraction methods tested have been carefully optimized so that the best performance for each method is reported. The smaller the SA values across the five minerals in Table 2.1, the better the results. It should be noted that Table 2.1 only displays the smallest SA scores of all endmembers with respect to each USGS signature for each algorithm. As a reference, the mean SA values across all five USGS signatures is also reported. The number of endmembers to be extracted was

Table 2.1 Spectral similarity scores (in degrees) between USGS mineral spectra and their corresponding endmembers extracted by several algorithms from the AVIRIS Cuprite scene

Algorithm	Alunite	Buddingtonite	Calcite	Kaolinite	Muscovite	Mean
	GDS84	GDS85	WS272	KGa-1	GDS107	
N-FINDR	4.81	4.29	7.60	9.92	5.05	6.33
OSP	4.81	4.16	9.52	10.76	5.29	6.91
VCA	6.91	5.38	9.53	9.65	6.47	7.59
MVSA	12.72	8.41	5.69	15.04	5.36	9.44
SISAL	9.78	5.13	12.78	13.53	8.00	9.84
AMEE	4.81	4.17	5.87	8.74	4.61	5.64
SSEE	4.81	4.16	8.48	10.73	4.63	6.57

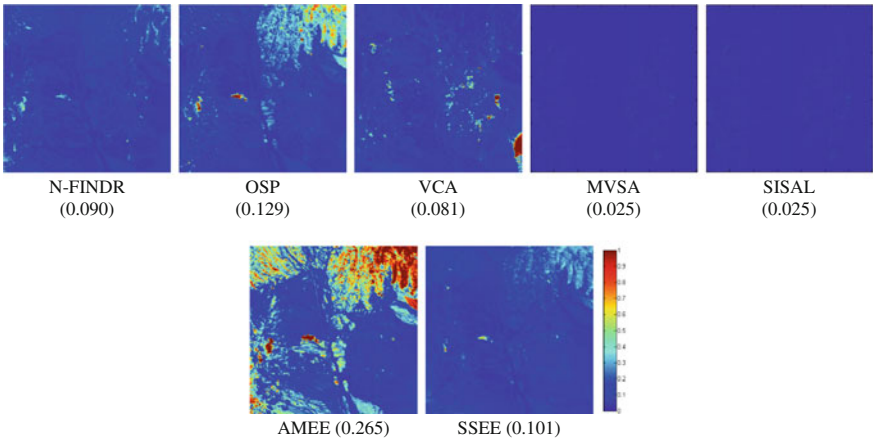


Fig. 2.14 Errors measured for various endmember extraction algorithms after reconstructing the AVIRIS Cuprite scene

set to $p = 19$ in all experiments after the consensus reached between HYSIME [11] and the VD concept [30], implemented using $P_F = 10^{-3}$ as the input false alarm probability. In this experiment, the best performance (in terms of SA) was obtained by the endmember extraction algorithms AMEE which include both spatial and spectral information.

Additionally, Fig. 2.14 shows the root mean squared error (RMSE) maps obtained after reconstructing the AVIRIS Cuprite scene using $p = 19$ endmembers extracted by different methods. As shown by this experiment, MVSA and SISAL provide the best results in terms of image reconstruction although they may provide unrealistic endmembers, as described in the previous experiment. To conclude this section, Table 2.2 reports the processing times of the compared algorithms.

Table 2.2 Processing times (in seconds) measured in a desktop PC with intel core i7 920 CPU at 2.67 Ghz with 4 GB of RAM

Algorithm	Total processing time
N-FINDR	466.08
OSP	136.09
VCA	31.12
MVSA	≃25000
SISAL	170.40
AMEE	76.06
SSEE	1051.23

2.3 Classification

The general hyperspectral image *classification* problem can be described as follows: At the input a B-band hyperspectral data cube is given, which can be considered as a set of n pixel vectors $\mathbf{X} = \{\mathbf{x}_j \in \mathbb{R}^B,\ j = 1, 2, \ldots, n\}$. Let $\Omega = \{\omega_1, \omega_2, \ldots, \omega_K\}$ be a set of information classes in the scene. Classification consists in assigning each pixel to one of the K classes of interest. An information class can represent either a physical substance (a ground cover material, for instance, *snow*, *water*, *wheat*), or a specific group of objects which may be made of several different physical materials (for instance, *roof*, *shadows*, *trees*).

In this chapter, we focus on *supervised* classification, which assumes that classes are defined by a set of training samples. Unsupervised classification, or clustering techniques have also been described in the literature. We refer the reader to [35] for a survey on unsupervised methods. An important assumption for classification techniques is that the spatial resolution of the image is high enough so that the data contains mostly pure pixels, i.e., pixels representing a single information class. In the opposite case, i.e., when the data is mostly composed of mixed pixels, spectral unmixing methods are more appropriate for image analysis.

The first attempts to classify hyperspectral images would assign each pixel to one of the classes based on its spectrum only [66]. These are often referred to as *pixelwise* (or non-contextual) classification techniques. However, with the increase of spatial resolution of hyperspectral sensors, objects in the image are typically large compared to the size of a pixel. In the ideal case, all the pixels of these objects should be assigned to the same class. It has then become then very important to simultaneously use spectral and spatial information for image classification [88]. *Spectral-spatial* classification (also referred to as spatial-contextual) assigns each pixel to one information class based on: (1) its own spectrum; (2) information extracted from its neighborhood, i.e., the spatial information. A multitude of methods have been proposed for this purpose, which differ in the ways of extracting spatial contextual information from the image scene and in the ways of combining spectral and spatial information.

The following sections review the keystone methods of pixelwise and spectral-spatial classification. We also include a detailed explanation of a recently-proposed mathematical model for spectral-spatial classification, based on a binary partition

tree representation [76]. This method first constructs a hierarchical region-based representation of the image stored in a tree structure, and then extracts objects of interest from the tree.

2.3.1 Supervised Pixelwise Classification

Landgrebe et al. were seminal in exploring procedures for hyperspectral data analysis and classification [66, 67]. They adapted pattern recognition procedures for this purpose. A simplified version of their proposed classification scheme, widely used until nowadays, is depicted in Fig. 2.15. There are two inputs to the system: the hyperspectral image and a set of observations of the ground which are labeled into classes of interest. From the hyperspectral image, there is first a process of feature extraction and selection. Features can be seen as an abstraction layer with meaningful descriptors derived from the raw input. This representation should me meaningful in the sense that it must be useful for the classification problem, describing and separating the classes of interest. Some input from the training labels themselves might be used to decide which features are relevant. The features associated to every pixel can be seen as a point in a high-dimensional space. The next step consists in *training* a classifier based on the set of labeled samples, i.e., partitioning the entire feature space into K exhaustive, nonoverlapping regions, so that every point in the feature space is uniquely associated with one of the K classes.

In the *pixelwise* approach, each image pixel is seen as a pattern to classify. One possibility is to use the pixel spectrum as the set of features that describe every pixel. Since this is often redundant, it is common to perform a more sophisticated feature extraction/selection step with the goal of reducing the dimensionality of the feature set and maximizing separability between classes. Different feature extraction techniques have been proposed and explored for this purpose, such as Discriminant Analysis Feature Extraction, Decision Boundary Feature Extraction and Nonparametric Weighted Feature Extraction [39, 67]. Once this step is accomplished, each pixel is classified according to its feature set.

The set of training samples is typically obtained by visually interpreting and manually labeling a small number of pixels in the data set, or by performing an *in situ* field campaign. The training data is used to define a model of the classes in the feature space. Assuming that each class can be described by a normal distribution,

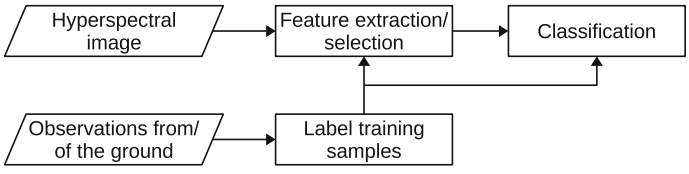


Fig. 2.15 Schematic diagram of the hyperspectral image classification process

Gaussian Maximum Likelihood classification has been for many years the standard thematic mapping procedure in hyperspectral remote sensing [99]. Essentially, it assigns a given pixel to the class ω_i that maximizes the posterior probability $P(\omega_i|\mathbf{x})$ in the Gaussian model. A serious drawback of this method consists in the primary assumption about the shape of the class-conditional probability density functions. If this assumption is wrong, classification results are no longer accurate. Furthermore, the high number of features available, usually coupled with a limited number of training samples, makes estimation of statistical class parameters unreliable. As a result, with a limited training set, the classification accuracy tends to decrease as the dimensionality increases, an issue often referred to as the *Hughes phenomenon* [66, 88]. High-dimensional spaces are mostly empty [62], making density estimation even more difficult.

In the 1990s, *neural network* approaches for classifying hyperspectral images received a lot of attention [7, 80, 111, 129]. Neural network models have an advantage over statistical methods in that they are distribution-free and thus no prior knowledge about the statistical distribution of classes is needed. In a neural network, a set of weighted sums and nonlinearities describe the function that classifies the input features. The training procedure involves finding the appropriate weights, which is done iteratively. The interest in such approaches greatly increased in the 1990s with improvements in the training techniques [6]. Yet there has been a limited use of neural networks for hyperspectral image classification primarily due to their algorithmic and training complexity [99]. Genetic algorithms for classification of hyperspectral data have also been presented [121], capable to deal with nonlinearly separable patterns but computationally demanding.

Early in this century, kernel methods such as Support Vector Machines (SVMs) have become very popular for hyperspectral image analysis, proving to be extremely well suited to classify high-dimensional data when a limited number of training samples is available [22, 48]. The SVM method seeks to trace an optimal hyperplane that linearly separates features into two groups with a maximum margin (see Fig. 2.16). A *soft margin* is typically used, where misclassified samples (i.e., on the wrong side of the hyperplane) are tolerated but penalized. To account for nonlinear separation boundaries, the data points are mapped to a higher-dimensional space by using a *kernel* function, and the linear SVM classification is performed on the transformed space. More details on SVMs can be found in Chap. 10. For hyperspectral image classification, two kernel functions have been widely used: the polynomial kernel and the Gaussian radial basis function. While initially devised for binary classification, the K-class problem can be solved by training K classifiers to distinguish each class from all the rest (*one vs all*) or $K(K-1)/2$ classifiers to distinguish every pair of classes (*one vs one*) [104].

To conclude, SVMs directly exploit the geometrical properties of data, without involving a density estimation procedure. This method can efficiently handle high-dimensional data, exhibiting low sensitivity to the Hughes phenomenon [59, 79]. Therefore, it is an excellent approach to attenuate the usually time-consuming feature extraction/selection procedure, thus simplifying the traditional pattern recognition scheme of Fig. 2.15. In hyperspectral image classification with SVMs, the dimen-

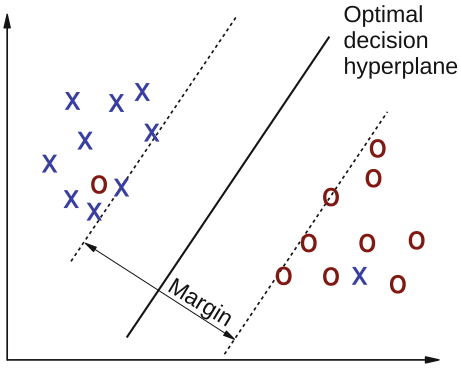


Fig. 2.16 Support vector machines (SVMs) search for an optimal hyperplane to linearly separate the data points with a maximum margin. This SVM uses a *soft* margin, adding robustness to difficult samples

sionality reduction step is often skipped and the spectrum directly used as the feature vector. Finally, SVMs exhibits a good generalization capability, fully exploiting the discrimination capability of the relatively few training samples available. All these advantages of the SVM method have made it the most widely used classifier for hyperspectral data in the last decade [13].

To further boost up classification accuracies, ensemble classification systems have been investigated for hyperspectral image classification. These approaches combine multiple learning algorithms to improve the predictive accuracy. Ham et al. [50] investigated the use of Random Forest framework and Ceamanos et al. [24] proposed an SVM-based ensemble approach, where separate SVM predictions are performed for subsets of spectral bands, and all outputs are used as the input for an additional SVM classifier.

All the described approaches assign each pixel to one of the classes based on its spectral properties alone, with no account being taken of how spatially adjacent pixels are classified. In the following, we summarize the key concepts for spectral-spatial classification of hyperspectral data.

2.3.2 Spectral-Spatial Classification

It is a proven fact that for images with high-spatial resolution, combining the spectral and the spatial information improves significantly the performance of classification methods. Surveys of spectral-spatial classification methods can be found in [13, 41].

A common spectral-spatial approach is to incorporate spatial information as part of the pixelwise classification process. Some feature extraction is applied to the surrounding area of a pixel and the result is integrated as part of the features associated to the individual pixel, in addition to the usual spectral features. In order to perform classification with a kernel method such as an SVM, the two sets of features must be combined. This can be done in different ways, ranging from a naive stacking of the feature vectors to more versatile methods. Different strategies of combining the two sources of information have been reviewed and compared in [23, 88].

To apply this scheme, one must define which is the neighborhood of a pixel from where spatial features are extracted. An idea as simple as the use of a fixed window already shows an improvement with respect to purely pixelwise approaches [88]. Benediktsson et al. [8] proposed to use morphological filters to obtain the spatial neighborhoods in an adaptive manner. In this method, a so-called structuring element is used to perform morphological opening and closing operations [109]. The effect of applying these operations is that image structures smaller than the structuring element are removed, otherwise preserved. These operations are applied with structuring elements of different sizes to create the *morphological profile*. This idea was applied in hyperspectral image classification [5] by computing the morphological profiles of the first principal components of the data, and combining them to obtain the features for classification.

Later on, Fauvel et al. [40] proposed to use the so-called self-complementary filters [110] for spatial feature extraction, which remove small structures from the image based on an area criterion, yielding a map of flat connected zones. This filter is applied on the first principal component of the hyperspectral image to extract adaptive spatial neighborhoods. The vector median [4] is then computed for each connected zone of the filtered result, and used as the spatial feature vector for all the pixels within the zone. Finally, SVM classification is performed with a weighted summation kernel to combine spectral and spatial information. More advanced morphological filters, called attribute filters, have been recently proposed to further enhance classification performance [3, 74].

Another important approach to characterize pixel entities using the spatial and the spectral information is the Markov random field (MRF) [42, 61]. MRFs (see also Chaps. 4 and 7) are probabilistic models widely used to include spatial context into image analysis schemes in terms of minimization of suitable energy functions [83]. The MRF energy function for image classification is commonly computed as a linear combination of a data term, which measures for each pixel the disagreement between a prior probabilistic model and the observed data, and a spatial context term, which expresses interaction between neighboring pixels. The first MRF-based models employed time-consuming energy minimization algorithms, such as iterated conditional modes and simulated annealing [82, 117]. More advanced methods, such as graph-cuts [19, 20] provided powerful alternatives from both theoretical and computational viewpoints, resulting in a growing use of the MRF-based models [72, 118]. For example, Tarabalka et al. [118] used probabilities derived from an SVM as the data term of an MRF energy, and used the α-expansion graph cut algorithm [20] to solve the K-class classification problem in hyperspectral imagery.

Finally, an important family of methods involves the segmentation of images and the classification of each of the individual segments. Segmentation methods partition an image into non-overlapping homogeneous regions with respect to some criterion of interest or homogeneity criterion (e.g., based on the intensity or on the texture) [46]. Hence, each region in the segmentation map can be seen as a connected spatial neighborhood for all the pixels within this region. One of the pioneering spatial-spectral techniques belongs to this category: the well-known ECHO (Extraction and Classification of Homogeneous Objects) classifier [65], which has been extensively used by

the remote sensing community. It is based on region growing to find homogeneous groups of adjacent pixels, which are then classified as single objects by a Gaussian maximum likelihood method. Since then, different techniques have been proposed for hyperspectral image segmentation, such as watershed, partitional clustering and Hierarchical Segmentation (HSeg) [112, 113, 116]. From a segmentation map, an SVM classifier and majority voting can be applied to combine spectral and spatial information: for every region in the segmentation map, all the pixels are assigned to the most frequent class within this region, based on SVM classification results [113]. This method yields an improvement of classification accuracies when compared to spectral-spatial techniques using local spatial neighborhoods.

It is however a challenging task to perform hyperspectral image segmentation automatically. The performance is highly dependent both on the measure of region homogeneity and on the algorithm parameters. Several alternatives have been proposed to deal with this challenge. Tarabalka et al. [114, 115] proposed to perform a marker-controlled segmentation for this purpose. The classification probabilities are used to automatically select the most reliably classified pixels (i.e., pixels belonging with the high probability to the assigned class). The classification map is then obtained by building a minimum spanning forest from the image graph rooted on the selected markers. Another alternative for automatic segmentation consists in building first a hierarchy of segmentations at different levels of details, and then selecting from this hierarchy the regions at different scales that correspond to the objects of interest. Valero et al. proposed to use a binary partition tree (BPT) model for this purpose [122]. In this method, a BPT is first constructed by iteratively clustering similar regions based on a criterion specifically designed for hyperspectral images. Each BPT node is then modeled by its mean spectrum and classified by using an SVM. A so-called misclassification rate is computed for each node, which can be understood as the error incurred by assigning the entire node to the wrong class. A spectral-spatial classification map is finally built in a bottom-up traversal of the tree by extracting regions with a low misclassification rate. In the next section we describe an energy minimization BPT-based model recently proposed in [76].

2.3.3 Object-Based Classification with Binary Partition Trees

The goal of classification is to convert the image data into tangible information that can be interpreted and incorporated into other systems. The ultimate elementary units which we want to identify are the *objects* present in the image. In the earlier years of remote sensing research, the per-pixel or sub-pixel analysis were particularly relevant given that pixel sizes were coarser than the objects themselves. The boundary between pixel-based and object-based analysis was still vague. As sensors improved their spatial resolution, objects started to be comprised of many pixels and object-based analysis emerged as a natural consequence of this. While pixelwise and spectral-spatial classification may constitute the first of a series of steps in the image analysis pipeline, object-based methods aim at delineating readily usable objects

from the image [15]. Contrary to the well-established pixelwise and spectral-spatial approaches described before, this section presents a recent object-based classification model for hyperspectral imagery, based on binary partition trees.

Binary partition trees (BPTs) were presented by Salembier and Garrido [103] as a way of representing a set of meaningful image regions in a compact and structured manner. The root node corresponds to the entire image, the following level represents the subdivision of the entire image into two disjoint regions, and so on. It constitutes then a hierarchical abstraction of an image, which can be navigated to extract meaningful regions at different scales. The typical workflow involves an initial tree construction stage, followed by a second stage of information extraction from the tree. For example, once a tree is constructed, an exhaustive segmentation of the image can be obtained by performing a horizontal "cut" on the structure (see Fig. 2.17). In this procedure, commonly referred to as *pruning*, branches can be selected at different scales, an inherent advantage of such hierarchical structure.

The construction of a BPT is done in a bottom-up fashion, by iteratively clustering pairs of similar regions together. The starting point is an initial subdivision of the image represented by a region adjacency graph (RAG), where every node conveys a region and the edges link spatial neighbors (i.e., candidates for merging). The typical initial RAG is the pixel grid, though nothing prevents the approach to be used with other inputs too (e.g., a RAG of small regions containing similar pixels, known as superpixel segmentation). Every edge in the RAG is labeled with a *dissimilarity* value that compares the two associated regions.

BPTs are constructed by following a global mutual best fitting region merging approach [69]: at each iteration, the two most similar regions in the current subdivision are merged together (i.e., the least weighted edge out of all edges in the RAG). When a merge occurs, a new region is added to the BPT, connected to its two corresponding children, as illustrated in Fig. 2.18). The process finishes when there are no more edges left in the RAG. A BPT constitutes then a record of the history of merges that occurred during the execution of a region merging algorithm.

The overall process can be implemented efficiently by using an updatable priority queue structure on top of the RAG edges to keep track of the highest priority element. Such a structure is first constructed in linear time and every subsequent update incurs in a logarithmic time cost. When two regions R_1 and R_2 are merged into a new

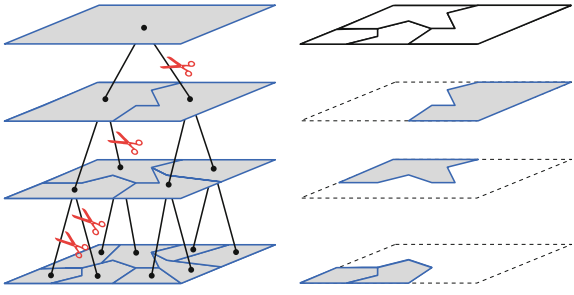


Fig. 2.17 A binary partition tree (BPT) is a hierarchical subdivision of an image. An exhaustive partitioning can be extracted by "cutting" branches at different scales

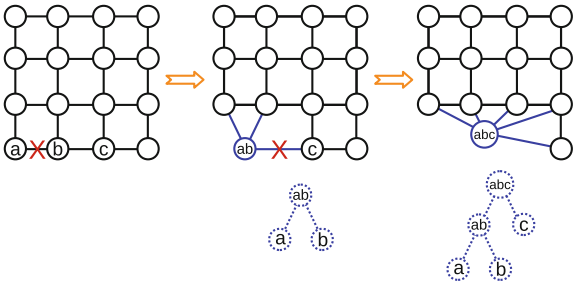


Fig. 2.18 A BPT is constructed by iteratively removing edges in a region adjacency graph (RAG). The resulting BPT encodes the history of the merges

region R_{12}, one must update the RAG (and the associated priority queue). The edge connecting R_1 and R_2 must be removed, but let us also remark that all the edges adjacent to R_1 and R_2 must also be eliminated from the RAG, since none of both regions exists anymore. We must then add the adjacency relations of the new region R_{12}. The computation is straightforward: the neighbors of the new region R_{12} are nothing but the union of the neighbors of the old R_1 and R_2 (with spatial care to remove any duplicates that may arise). The dissimilarity value associated to each of these edges must be computed and pushed to the priority queue. The complexity of the overall BPT construction process is $O(n \log(n) M)$, n being the initial number of nodes and M the maximum number of neighbors of a merged region during the construction. Given that typically $M \ll n$, the algorithm is quasilinear in practice.

The final tree contains exactly $2n - 1$ nodes, which is a very space-efficient representation. Let us remark though that only a subset of all possible planar subdivisions is represented by the tree, hence the research efforts to construct a good initial tree that conveys meaningful objects of the underlying image.

The key elements to define the behavior of a BPT are the *region model*, i.e. how regions are represented, and the *dissimilarity function*, i.e., the function to compare the region models, used to define the priority of the merges during tree construction. The next paragraphs review the contributions related to these two elements.

Region Model

The object-based nature of BPTs allows to have rich representations of the region that go beyond pixel spectra. Every BPT node can convey *regional* information, describing the region as a whole and not as a set of individual pixels. The standard variation of the spectral signatures in the region or shape features such as compactness are some of the regional data that can be associated to every node.

To represent the spectrum of a region (and then compare it to the spectra of other regions) there are essentially two alternatives: parametric and non-parametric models. A parametric model makes assumptions about the homogeneity or Gaussian distribution inside the regions. A typical parametric model is to represent the spectrum of a region as the mean spectrum of its pixels. Non-parametric models, on the contrary, consist of per-band histograms of the pixel values, hence they represent the real observed distributions. In hyperspectral imagery, non-parametric models have a better performance since they can describe the internal variability of a region [122].

For example, a texture might correspond to several peaks in the histogram. When averaging spectra in regions with high variability, one might end up representing the region with a "false" spectrum that is not present in any of the individual pixels.

In addition to spectral data, the model usually stores the *area* of the region, since it is commonly used in the dissimilarity function. Other shape descriptors such as *solidity*, *rectangularity index*, *elongatedness* and *compactness* can also be efficiently stored and computed from the children nodes [77].

Dissimilarity Function

To establish a priority for merging during BPT construction, it is required to provide a means to compare models of two regions. A dissimilarity function $O(R_1, R_2)$ typically used for this purpose comprises two factors as follows:

$$O(R_1, R_2) = \min(|R_1|, |R_2|)^{\beta} D(R_1, R_2), \tag{2.5}$$

where $|R_i|$ denotes the area of region R_i. The first part of (2.5), $\min(|R_1|, |R_2|)^{\beta}$, is the so-called *area-weighting* factor. This is an agglomerative force intended to cluster regions that are very small compared to the rest of the elements in the RAG. When no area-weighting is used (i.e., $\beta = 0$), the resulting BPT might isolate small noisy areas and connect them to the rest only near the root of the tree. With moderate values of β, small regions are merged at some point, forcing the trees to better look like a hierarchical subdivision. When β is too large, the trees might be too biased to be balanced, hampering their representation capabilities. Even though this parameter is barely discussed in the literature, being mostly set to $\beta = 0.5$ or $\beta = 1$, we must point out that it is indeed a parameter that has to be selected. In our experience, no area-weighting leads to poor representations (e.g., the root containing two children: one noisy pixel and all the rest of the image), while low values of β solve this issue without biasing the trees too much. Alternatively, Calderero and Marques [21] proposed to keep track of the out-of-scale regions and force their merging at some point, while Valero et al. [122] used a weighted sum of pixel values in a window to initialize the histograms, as a way of smoothing out outliers.

The second factor, $D(R_1, R_2)$, compares both regions based on their spectra. Kullback-Leiber divergence and Bhattacharyya distance are popular choices both in hyperspectral imagery and other types of images [21, 122]. Spectra are seen as probability distributions and compared using standard information theory concepts. Every bin of one histogram is compared against the corresponding bin of the other histogram. However, using cross-bin measures, which go beyond individual bins, has proven to be more robust [122]. The average of Earth Mover's Distances [101] among histograms of all bands can be used as a robust and efficient cross-bin dissimilarity function. Every distribution is seen as a pile of dirt, and the difference between two distributions is seen as the amount of work required to turn one pile into the other one.

2.3.3.1 Multi-class Segmentation with BPTs

The problem of object-based classification can be seen as the simultaneous segmentation of an image and the assignment of a label to every segment. This section first formulates this problem as the minimization of an energy, and describes an algorithm to extract the optimal segmentation with respect to that energy from a BPT. This algorithm outputs the lowest-energy solution from all the segmentations represented by the BPT, which are a subset of all possible image partitions. It is then a matter of high importance to construct good BPTs whose solution space contains relevant candidates for object-based analysis. For this we describe a supervised BPT construction technique that incorporates class probabilities to cluster objects together.

Multi-class Segmentation as Energy Minimization

Let $\mathbf{X} = \{\mathbf{x}_i \in \mathbb{R}^B, i = 1, 2, \ldots, n\}$ be a B-band image seen as a set of n pixel vectors. Multi-class segmentation consists in an exhaustive partitioning of the pixels into a non-overlapping set of regions $\mathrm{R} = (R_j)$, with associated labels $\mathrm{L} = (L_j)$, where every label L_j belongs to the set Ω of available information classes. From each object class, we suppose we are given training examples from which we can derive posterior probabilities $P(L_j|\mathbf{x}_i)$ of assigning a certain label L_j after the spectral observation $\mathbf{x}_i$ is taken into account. Such posterior probability may be derived from a support vector machine [128]. The negative log-likelihood $-\log P(L_j|\mathbf{x}_i)$ is typically used to express a *cost* that penalizes the assignment of label L_j to pixel $\mathbf{x}_i$.

Our task is to find the labeled partitioning (R, L) from a BPT that minimizes the following energy:

$$E(\mathrm{R}, \mathrm{L}) = \lambda||\mathrm{R}|| - \sum_{R_j \in \mathrm{R}} \sum_{\mathbf{x}_i \in R_j} \log P(L_j|\mathbf{x}_i). \tag{2.6}$$

Let us first observe that the same label L_j is assigned to all pixels $\mathbf{x}_i$ in region R_j, since the entire segments take a single label. The first term is a regularizer on the number of regions in the partition $||\mathrm{R}||$, and controls the coarseness of the output through parameter λ. In the absence of this term (i.e., $\lambda = 0$), the optimal solution is to create one segment per pixel and assign to it the lowest-cost label. To introduce the notion of object we must then set $\lambda > 0$. We here set this parameter manually, but let us mention that in recent work the regularization term was directly learned from training samples [77].

From a BPT, the best possible labeled segmentation with respect to Eq. 2.6 can be extracted efficiently [102]. This task can be interpreted as the extraction of a minimal horizontal s-t *cut* on the tree (see Fig. 2.17), i.e., with a source at every leaf and a sink at the root. Let us denote $C(R)$ the energy of the cut on R with minimal (2.6) among all possible cuts.

Considering that the branches in the tree are independent, the globally optimal cut can be found by a dynamic programming algorithm. Let us denote $\mathscr{E}(R) = \min_{L \in \Omega} E(\{R\}, \{L\})$ the lowest possible energy of a region R (by assigning the label

that incurs the lowest cost). The tree is traversed in a bottom-up manner. Whenever a region R is visited, the following property is evaluated:

$$\mathscr{E}(R) \leqslant C(R_{left}) + C(R_{right}), \tag{2.7}$$

where R_{left} and R_{right} are the children of R. If the property does not stand, we set $C(R) = C(R_{left}) + C(R_{right})$ and keep the best cuts of both children. Otherwise, we set $C(R) = \mathscr{E}(R)$ and replace the cuts by R with label L. This process is executed recursively until reaching the root of the tree. The overall algorithm is linear in the image size, since only one BPT traversal is required, and guarantees the optimal cut in the space of solutions represented by the BPT.

Supervised BPT Construction

Even though the globally optimal cut on a BPT can be found efficiently, not all the possible ways of segmenting an image are represented in the structure. In some images, objects have considerable internal variability. For example, it is known that the different parts of a roof often contrast more with each other than with other surrounding objects [45]. This is more prevalent in high-resolution imagery and in cluttered urban scenes.

In such images, it is common to observe objects that are split into different branches of the tree instead of being contained in a single node. This behavior is illustrated in Fig. 2.19, which shows a BPT built on an image of a non-uniform roof. During BPT construction, a part of the roof (b) is merged first to something else (a) than to the rest of the object (c-d-e) because it is more similar in terms of Eq. (2.5). As a consequence, the entire building (b-c-d-e) cannot be extracted by selecting a single node in the tree.

Figure 2.20 illustrates this phenomenon on real image data. Two fragments of the *Pavia Center* image, which will be introduced in the experimental section, are shown in Fig. 2.20a. The scene contains multiple buildings, streets and cars adjacent to each other. A BPT with a non-parametric region model and Earth Mover's Distance was constructed for this image. The energy minimization scheme (2.6) was then applied, and the objects labeled as *tile* isolated from the rest to aid the interpretation. Figure 2.20b depicts the surface covered by *tiles*, as predicted by the BPT cut. This way of illustrating the classification is purely pixelwise, since no distinction about the *objects* extracted from the tree is made. This is a common way of illustrating results in the literature, even when the goal is to perform object detection (e.g., [123]).

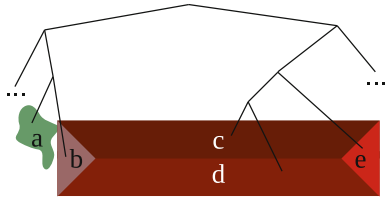


Fig. 2.19 Faulty BPT: the object (*bcde*) is not represented in a single node, since a part of it (*b*) merged first to something else

However, observing the actual objects extracted from the BPT (Fig. 2.20c) we can see that the regions hardly correspond to actual objects in the image. Even though the surface covered by these objects might be satisfactory from a pixelwise perspective, an object-based analysis would certainly be less impressive.

Let us recall that the use of non-parametric region models is to represent internal variability. However, commonly used dissimilarity functions such as (2.5) penalize the merging of dissimilar regions. In an unsupervised context, where there is no notion of object class, there is little to do to deal with this, since there is no reason to cluster dissimilar regions together. However, when class probabilities are available we propose to include an additional force that clusters regions belonging to the same class, despite being spectrally dissimilar. The new function is as follows:

$$O(R_1, R_2) = \min(|R_1|, |R_2|)^{\beta}\Big[(1-\alpha)D(R_1, R_2) - \alpha \log P(L_{R_1} = L_{R_2})\Big]. \tag{2.8}$$

As in the original dissimilarity function (2.5), there is an area-weighting factor and an unsupervised term $D(R_1, R_2)$, which is computed by comparing spectral histograms of regions without any preliminary training. Equation 2.8 adds a *supervised* term $P(L_{R_1} = L_{R_2}|R_1, R_2)$, the probability of assigning the same label to both regions. This way, while the unsupervised term penalizes spectral dissimilarity, the supervised term will encourage merging regions that are likely to belong to the same class. The trade-off between both terms is controlled by parameter α.

The term $P(L_{R_1} = L_{R_2}|R_1, R_2)$ is computed by marginalizing over the classes as follows:

$$P(L_{R_1} = L_{R_2}|R_1, R_2) = \sum_{j=1}^{K} P(L_j|R_1)P(L_j|R_2), \tag{2.9}$$

(a) Color composition (b) Areas labeled *tile* (c) Extracted objects

Fig. 2.20 Analyzing classification from a pixelwise **b** vs object-based **c** perspective. Even though the area covered by the *tile* objects might be satisfactory (**b**), the objects that constitute this area often do not correspond to the real objects (**c**)

where K is the number of classes and $P(L_j|R_k)$, with $k \in \{1, 2\}$, represents the probability of assigning a certain label L_j to segment R_k. We must now define a way to compute $P(L_j|R_k)$ based on the posteriors of the individual pixels contained in the region. One way to do this is to compute the probability of assigning the label to all pixels, conditioned by the fact that all labels are known to be equal inside the region:

$$P(L_j|R_k) = \prod_{\mathbf{x}_i \in R_k} P(L_j|\mathbf{x}_i) / \left[\sum_{\omega_m \in \Omega} \prod_{\mathbf{x}_i \in R_k} P(\omega_m|\mathbf{x}_i) \right]. \tag{2.10}$$

Alternatively, one can estimate $P(L_j|R_k)$ by averaging the individual pixel probabilities:

$$P(L_j|R_k) = \frac{1}{|R_k|} \sum_{\mathbf{x}_i \in R_k} P(L_j|\mathbf{x}_i). \tag{2.11}$$

While the first expression is closer to a strict Bayesian interpretation, we found the second one to be a simple yet useful approximation.

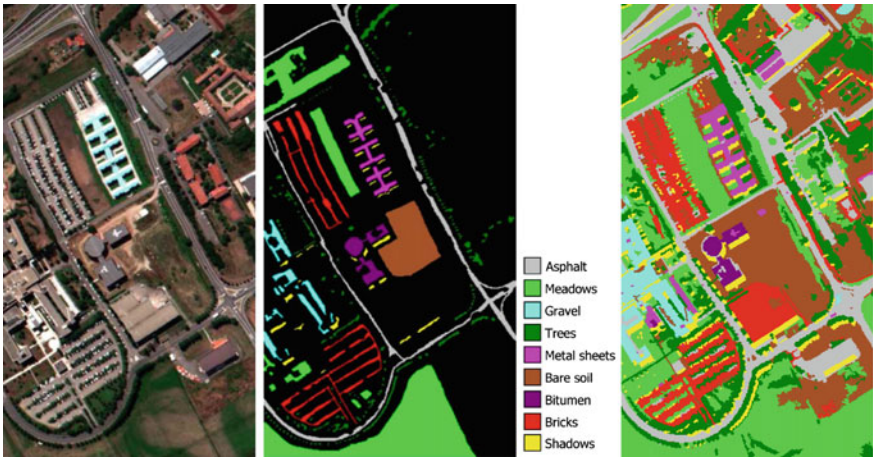
By introducing (2.8) we expect to better cluster semantically significant objects together. The advantage of such an outcome is two-fold: first of all, the classification accuracy is improved. Secondly, there is a notion of object, which constitutes a higher-level interpretation of the input image rather than mere pixelwise labeling.

2.3.4 Experimental Results

This sections describes two series of experiments to analyze and compare different methods of hyperspectral image classification. We report results for the most representative pixelwise and spectral-spatial methods discussed in the previous sections, as well as the BPT model. The first set of experiments is performed on a dataset over the *University of Pavia*, Italy. The goal of this evaluation is to compare the different approaches in terms of per-pixel classification accuracy, with the particular goal of verifying that the introduction of spatial information improves the results.

A second set of experiments is carried out on the *Pavia center* hyperspectral dataset. The goal of these experiments is to evaluate the behavior of the techniques from an object-based perspective, providing an object overlap measure between reference and detected objects. We compare the typical unsupervised BPT construction approach and the supervised alternative introduced in Sect. 2.3.3.1.

Both images were acquired with the Reflective Optics System Imaging Spectrometer (ROSIS-03). This optical sensor provides 115 bands with a spectral coverage ranging from 0.43 to 0.86 μm and 1.3 m spatial resolution.



(a) Color composition (b) Reference data (c) $\text{BPT}_{\alpha=0.5}$ classification

Fig. 2.21 *University of Pavia* hyperspectral dataset

University of Pavia

This image is of size 610×340 and contains 103 spectral channels (after excluding 12 noisy bands). Figure 2.21a illustrates a false color composition of the hyperspectral image. The reference data contains nine classes of interest, as depicted in Fig. 2.21b.

In the next paragraphs we summarize the classification methods that were executed and compared on this dataset. As described in Sects. 2.3.1–2.3.3, these techniques have become the standard in the remote sensing literature.

SVM A support vector machine (SVM) was trained on 50 randomly selected samples for every class. A multi-class one vs one SVM with Gaussian kernel was used (with parameters $C = 128$ and $\gamma = 0.125$, set by fivefold cross-validation).
Graph cut A graph cut with α-expansion [20] (which proved to be effective in hyperspectral image classification [118]) was executed on probabilities derived the SVM. Its regularity parameter was set empirically to optimize the accuracy.
HSeg The technique presented in [113] was also implemented, which consists in first performing a segmentation and then labeling every segment. A recursive hierarchical image segmentation (HSeg) is used, followed by a majority voting procedure in which every segment is labeled as the majority class of the SVM predictions inside the segment. Parameter 'spclust_wght' was set to 0.1, following the original publication.
BPT For the binary partition tree (BPT) model described in this chapter, the tree was constructed by using a non-parametric model with 30 histogram bins per band, the Earth Mover's Distance to compare histograms and mild area-weighting ($\beta = 0.1$). The coarseness parameter λ in (2.6) was empirically set to 40. Two variants were tested: (**a**) totally unsupervised construction, i.e., setting $\alpha = 0$ in (2.8), which is equivalent to the old function (2.5); (**b**) supervised construction with equal contribution from both terms in (2.8), i.e., $\alpha = 0.5$.

Table 2.3 Numerical evaluation on *University of Pavia* dataset (in %)

	SVM	Graph cut	HSeg [113]	$BPT_{\alpha=0}$	$BPT_{\alpha=0.5}$
AA	88.03	95.13	95.35	95.49	**97.32**
OA	80.38	91.69	90.75	**94.45**	93.13
Asphalt	77.66	94.58	95.40	97.83	**99.15**
Meadows	72.74	86.10	83.63	**92.65**	86.07
Gravel	79.55	86.58	98.98	84.87	**99.17**
Trees	95.95	**97.38**	96.28	91.44	96.72
Metal	99.61	**100.0**	99.15	99.07	99.92
Bare soil	89.38	**98.39**	94.86	97.99	98.31
Bitumen	94.37	95.55	95.23	**99.92**	97.19
Bricks	82.89	97.74	97.88	95.59	**99.31**
Shadows	**100.0**	99.89	96.77	**100.0**	**100.0**

The test dataset was created by excluding the pixel used from SVM from the ground truth. To measure the performance we use the average accuracy (AA) and overall accuracy (OA). The first one computes for every class the percentage of correctly classified pixels from the test data, and averages these values over all the classes. The latter is the proportion of correctly classified pixels. The pixels used for SVM training are excluded in the evaluation. The numerical results are deployed on Table 2.3. The accuracies for individual classes are also included in the table. We can verify that purely pixelwise methods such as SVM have a lower performance than spectral-spatial approaches. The BPT models (with $\alpha = 0$ and $\alpha = 0.5$) outperform the other techniques. The inclusion of class probabilities in tree construction ($\alpha = 0.5$) boosts the AA with a mild decrease of OA with respect to the unsupervised construction.

The overall classification map for the $BPT_{\alpha=0.5}$ method is shown in Fig. 2.21c and two fragments are amplified and compared with other methods in Fig. 2.22. These results show that in general BPTs constitute an improvement with respect to the other techniques. The benefit of supervised ($\alpha = 0.5$) over unsupervised ($\alpha = 0$) construction is not entirely clear in this dataset. First of all, there are few objects of every class in the reference data and the labeled pixels do not cover the entire surface of the objects. Moreover, there seems to be a significant contrast between objects and their surroundings, a situation in which the supervised term in (2.8) may not be very relevant. While BPTs have proved to be competitive from a pixelwise perspective, we require a different dataset to evaluate the performance of the methods from an object-based perspective and compare the unsupervised vs supervised construction models.

Pavia Center

This image has spatial dimensions 400×300 and contains 102 bands. A color composition of the image is shown in Fig. 2.23a. Compared to the *University* dataset, this

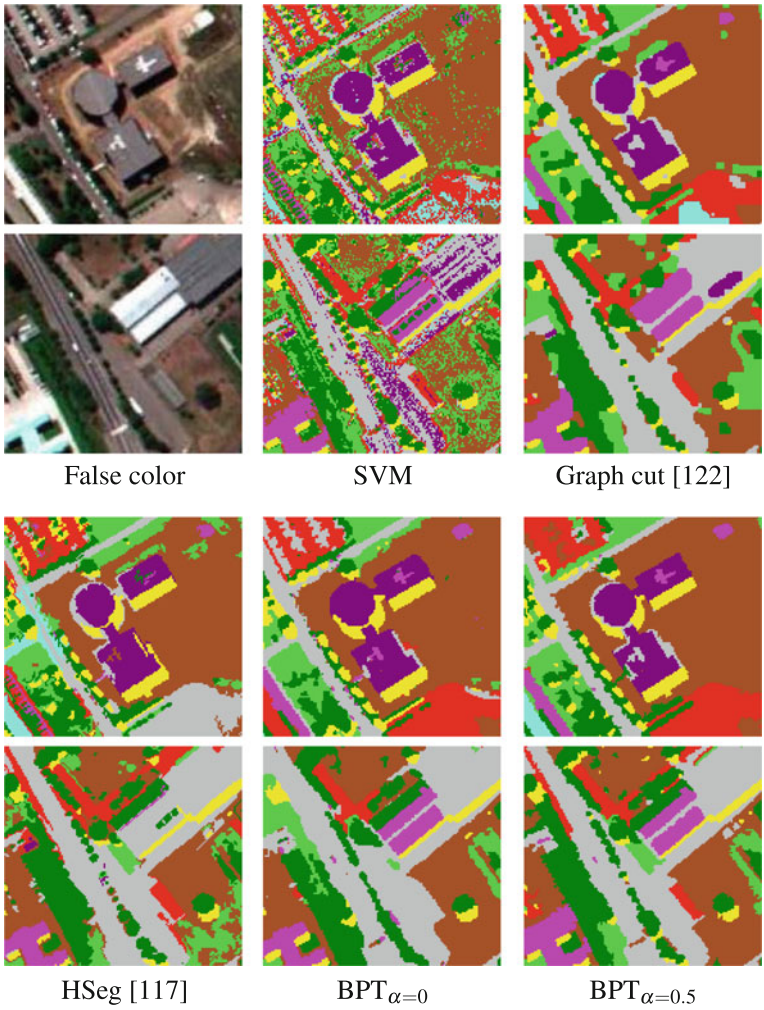
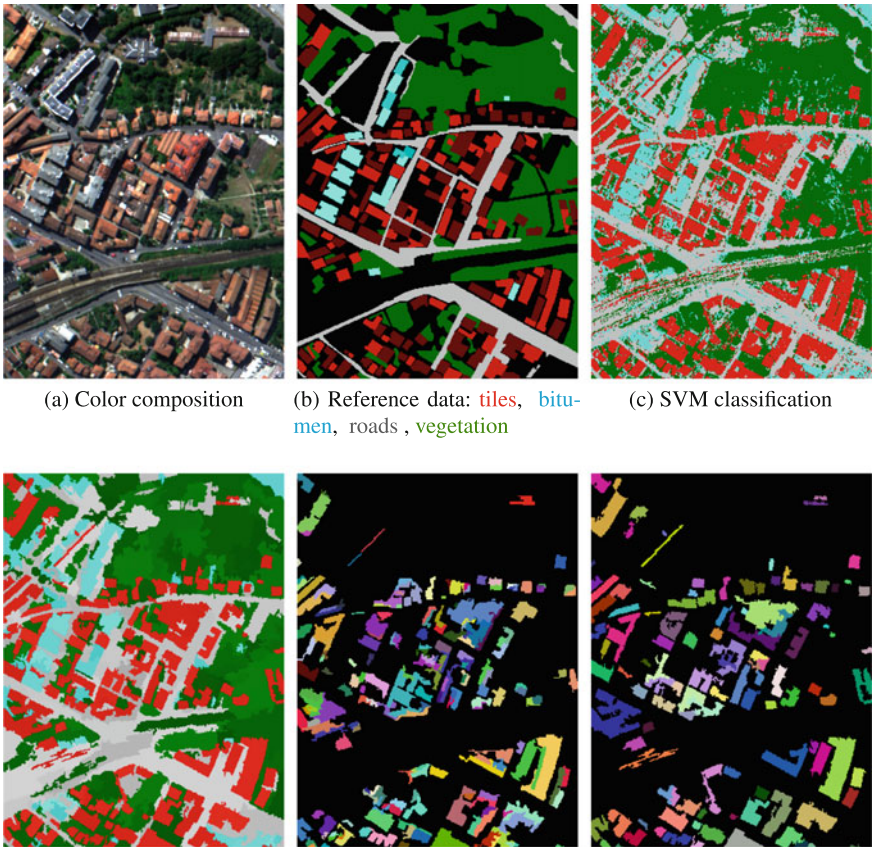


Fig. 2.22 *University of Pavia*: closeups of classification maps with different methods

image presents a more cluttered scene with objects composed of dissimilar parts. As illustrated previously in Fig. 2.20, BPT nodes may not correspond to entire objects because parts of them grow into other adjacent objects during the construction.

A reference image that labels entire objects and not just isolated pixels was built, including four classes (see Fig. 2.23b). This reference was constructed by combining the labeling of isolated pixels provided with the original image, visual inspection and official Italian records of building boundaries, which are available for this area through the OpenStreetMap.org database. Since the boundaries of buildings are well defined, there is a particular interest in analyzing the performance of BPTs to extract buildings.



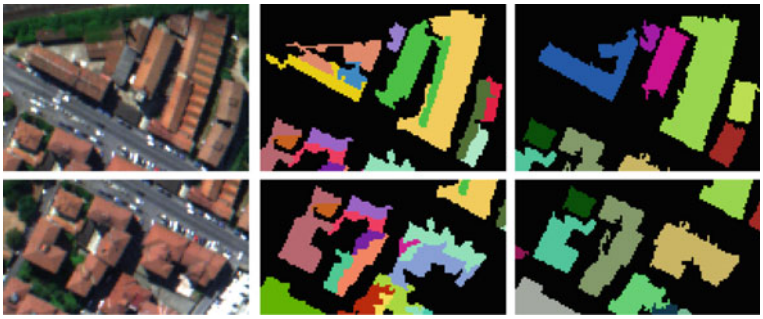
(a) Color composition (b) Reference data: tiles, bitumen, roads , vegetation (c) SVM classification

(d) BPT classification ($\alpha = 0.5$) (e) Tile building detection (BPT with $\alpha = 0$) (f) Tile building detection (BPT with $\alpha = 0.5$)

Fig. 2.23 Experiments on *Pavia Center* hyperspectral image

An SVM is first trained on randomly selected samples (parameters $C = 128$, $\gamma = 2^{-5}$). The SVM classification is shown in Fig. 2.23c. A BPT is then constructed on top of the SVM probabilities, in a similar experimental setting as with the *University of Pavia* dataset, and the classification map is extracted by setting $\lambda = 20$ in Eq. (2.6). The resulting classification map with supervised tree construction ($\alpha = 0.5$) is shown in Fig. 2.23d.

Figure 2.23e, f and the close-ups of Fig. 2.24 compare the results obtained by applying the unsupervised and supervised approaches for BPT construction. These figures isolate the *tile* objects from the rest and assign a random color to every individual object. From these illustrations we can appreciate that including class probabilities during BPT construction has the effect of better clustering the objects together. To validate this numerically we compute the overlap between every building (belonging either to *tiles* or *bitumen* classes) in the reference data and the most



(a) Color composition (b) BPT with $\alpha = 0$ (c) BPT with $\alpha = 0.5$

Fig. 2.24 Unsupervised **b** versus supervised **c** BPT construction. In the supervised case, regions are better clustered together to represent significant objects

Table 2.4 Numerical evaluation on the *Pavia Center* dataset

	SVM	Graph cut	$BPT_{\alpha=0}$	$BPT_{\alpha=0.5}$
Building overlap	0.51	0.51	0.54	**0.56**
Overall accuracy	0.88	**0.94**	0.91	**0.94**

overlapping building region in the BPT output. The overlap is measured with Dice's coefficient defined as: $2|R_1 \cap R_2|/(|R_1| + |R_2|)$. The resulting overlap coefficients are averaged over all reference buildings to produce an estimation of how well the BPT output matches the reference data from an object-based perspective. The numerical results, together with the overall accuracy, are summarized in Table 2.4, which also includes the values for SVM and graph cut. A first observation we can make is that $BPT_{\alpha=0.5}$ performs better than $BPT_{\alpha=0}$, corroborating the visual impression from Fig. 2.23e, f. Secondly, while graph cut is known to improve the SVM classification, we can see that this is true from a pixelwise perspective (in terms of OA) but not from an object-based perspective (in terms of building overlap). Finally, the use of $BPT_{\alpha=0.5}$ outperforms the other methods in terms of object overlap. This validates the idea of including class probabilities during tree construction for a better object-based analysis of hyperspectral imagery.

2.4 Challenges

The classification of hyperspectral imagery presents a number of challenges proper to the nature of this image modality. The integration of spatial and spectral information is one of the most widely addressed issues, as we have reviewed throughout this chapter. This concern will certainly continue to intrigue the scientific community and will remain an active research area. However, the imbalance between the high

dimensionality of hyperspectral data and the low amount of training samples is still probably one of the largest sources of difficulty.

There is a growing trend to study new training schemes to deal with the limited availability of labeled data. Notably, *semi-supervised* algorithms are arising in the hyperspectral literature (e.g., [97, 119]). These algorithms combine a low amount of labeled training data with *unlabeled* samples, under the assumption that the latter can be obtained with little effort. A smart combination of labeled and unlabeled data may significantly improve the accuracy of classification. Among semi-supervised algorithms, *active learning* methods interact with the user to actively query for helpful labels [86, 120, 125].

With the recent advent of *deep learning* in multiple application domains, it will certainly gain increasing attention in the hyperspectral image analysis community. Some first research efforts in this direction can be already identified in the literature [34, 75].

To conclude, we can say that hyperspectral remote sensing image analysis uses and adapts frontier concepts, frameworks and algorithms from the fields of signal and image processing, statistical inference and machine learning. The compendium of techniques presented in this chapter reflects the increasing sophistication of a field that is rapidly maturing at the intersection of many different disciplines.

Acknowledgements This work has been supported by:
- Junta de Extremadura (decreto 297/2014, ayudas para la realización de actividades de investigación y desarrollo tecnológico, de divulgación y de transferencia de conocimiento por los Grupos de Investigación de Extremadura, Ref. GR15005);
- Centre national d'études spatiales (CNES, France).

References

1. Adams, J.B., Smith, M.O., Johnson, P.E.: Spectral mixture modeling: a new analysis of rock and soil types at the Viking lander 1 site. J. Geophys. Res. **91**, 8098–8112 (1986)
2. Ambikapathi, A., Chan, T.-H., Ma, W.-K., Chi, C.-Y.: Chance-constrained robust minimum-volume enclosing simplex algorithm for hyperspectral unmixing. IEEE Trans. Geosci. Remote Sens. **49**(11), 4194–4209 (2011)
3. Aptoula, E., Dalla Mura, M., Lefèvre, S.: Vector attribute profiles for hyperspectral image classification. IEEE Trans. Geosci. Remote Sens. **54**(6), 3208–3220 (2016)
4. Astola, J., Haavisto, P., Neuvo, Y.: Vector median filters. Proc. IEEE **78**(4), 678–689 (1990)
5. Benediktsson, J.A., Palmason, J.A., Sveinsson, J.R.: Classification of hyperspectral data from urban areas based on extended morphological profiles. IEEE Trans. Geosci. Remote Sens. **43**(3), 480–491 (2005)
6. Benediktsson, J.A., Swain, P.H.: Statistical Methods and Neural Network Approaches for Classification of Data from Multiple Sources. Ph.D. thesis, Purdue Univ., School of Elect. Eng., West Lafayette, IN (1990)
7. Benediktsson, J.A., Swain, P.H., Ersoy, O.K.: Conjugate gradient neural networks in classification of very high dimensional remote sensing data. Int. J. Remote Sens. **14**(15), 2883–2903 (1993)

8. Benediktsson, J.A., Pesaresi, M., Arnason, K.: Classification and feature extraction for remote sensing images from urban areas based on morphological transformations. IEEE Trans. Geosci. Remote Sens. **41**(9), 1940–1949 (2003)
9. Berman, M., Kiiveri, H., Lagerstrom, R., Ernst, A., Dunne, R., Huntington, J.F.: ICE: a statistical approach to identifying endmembers in hyperspectral images. IEEE Trans. Geosci. Remote Sens. **42**(10), 2085–2095 (2004)
10. Bioucas-Dias, J.: A variable splitting augmented lagragian approach to linear spectral unmixing. In: IEEE Whispers, pp. 1–4 (2009)
11. Bioucas-Dias, J.M., Nascimento, J.M.P.: Hyperspectral subspace identification. IEEE Trans. Geosci. Remote Sens. **46**(8), 2435–2445 (2008)
12. Bioucas-Dias, J.M., Plaza, A.: Hyperspectral unmixing: geometrical, statistical, and sparse regression-based approaches. In: Proceedings of the SPIE Image and Signal Process, Remote Sens, XVI, vol. 7830, pp. 1–15 (2010)
13. Bioucas-Dias, J.M., Plaza, A., Camps-Valls, G., Scheunders, P., Nasrabadi, N.M., Chanussot, J.: Hyperspectral remote sensing data analysis and future challenges. IEEE Geosci. Remote Sens. Mag. **1**(2), 6–36 (2013)
14. Bioucas-Dias, J.M., Plaza, A., Dobigeon, N., Parente, M., Du, Q., Gader, P., Chanussot, J.: Hyperspectral unmixing overview: geometrical, statistical and sparse regression-based approaches. IEEE J. Sel. Topics Appl. Earth Obs. Remote Sens. **5**(2), 354–379 (2012)
15. Blaschke, T.: Object based image analysis for remote sensing. ISPRS J. Photogr. Remote Sens. **65**(1), 2–16 (2010)
16. Boardman, J.W., Kruse, F.A., Green, R.O.: Mapping target signatures via partial unmixing of Aviris data. Proc. JPL Airborne Earth Sci. Workshop **95–7**, 23–26 (1995)
17. Boggs, J.L., Tsegaye, T.D., Coleman, T.L., Reddy, K.C., Fahsi, A.: Relationship between hyperspectral reflectance, soil nitrate-nitrogen, cotton leaf chlorophyll, and cotton yield: a step toward precision agriculture. J. Sustain. Agric. **22**(3), 5–16 (2003)
18. Borel, C.C., Gerslt, S.A.W.: Nonlinear spectral mixing models for vegetative and soil surfaces. Remote Sens. Environ. **47**, 403–416 (1994)
19. Boykov, Y., Kolmogorov, V.: An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. IEEE Trans. Pattern Anal. Mach. Intell. **26**(9), 1124–1137 (2004)
20. Boykov, Y., Veksler, O., Zabih, R.: Fast approximate energy minimization via graph cuts. IEEE Trans. Pattern Anal. Mach. Intell. **23**(11), 1222–1239 (2001)
21. Calderero, F., Marques, F.: Region merging techniques using information theory statistical measures. IEEE Trans. Image Process. **19**(6), 1567–1586 (2010)
22. Camps-Valls, G., Bruzzone, L.: Kernel-based methods for hyperspectral image classification. IEEE Trans. Geosci. Remote Sens. **43**(6), 1351–1362 (2005)
23. Camps-Valls, G., Gomez-Chova, L., Muñoz-Marí, J., Vila-Francés, J., Calpe-Maravilla, J.: Composite kernels for hyperspectral image classification. IEEE Geosci. Remote Sens. Lett. **3**(1), 93–97 (2006)
24. Ceamanos, X., Waske, B., Benediktsson, J.A., Chanussot, J., Fauvel, M., Sveinsson, J.R.: A classifier ensemble based on fusion of support vector machines for classifying hyperspectral data. Intern. J. Image Data Fusion **1**(4), 293–307 (2010)
25. Chan, T.-H., Chi, C.-Y., Huang, Y.-M., Ma, W.-K.: A convex analysis based minimum-volume enclosing simplex algorithm for hyperspectral unmixing. IEEE Trans. Signal Process. **57**(11), 4418–4432 (2009)
26. Chan, T.-H., Ma, W.-K., Ambikapathi, A., Chi, C.-Y.: A simplex volume maximization framework for hyperspectral endmember extraction. IEEE Trans. Geosci. Remote Sens. **49**(11), 4177–4193 (2011)
27. Chang, C., Zhao, X., Althouse, M.L.G., Pan, J.J.: Least squares subspace projection approach to mixed pixel classification for hyperspectral images. IEEE Trans. Geosci. Remote Sens. **36**(3), 898–912 (1998)
28. Chang, C.-I.: Hyperspectral Imaging: Techniques for Spectral Detection and Classification. Kluwer Academic/Plenum Publishers, New York (2003)

29. Chang, C.-I.: Hyperspectral Data Exploitation: Theory and Applications. Wiley, New York (2007)
30. Chang, C.-I., Du, Q.: Estimation of number of spectrally distinct signal sources in hyperspectral imagery. IEEE Trans. Geosci. Remote Sens. **42**(3), 608–619 (2004)
31. Chang, C.-I., Heinz, D.: Constrained subpixel target detection for remotely sensed imagery. IEEE Trans. Geosci. Remote Sens. **38**, 1144–1159 (2000)
32. Chang, C.-I., Plaza, A.: A fast iterative algorithm for implementation of pixel purity index. IEEE Geosci. Remote Sens. Lett. **3**(1), 63–67 (2006)
33. Chang, C.-I., Wu, C.-C., Liu, W., Ouyang, Y.-C.: A new growing method for simplex-based endmember extraction algorithm. IEEE Trans. Geosci. Remote Sens. **44**(10), 2804–2819 (2006)
34. Chen, Y., Zhao, X., Jia, X.: Spectral-spatial classification of hyperspectral data based on deep belief network. IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens. **8**(6), 2381–2392 (2015)
35. Chutia, D., Bhattacharyya, D.K., Sarma, K.K., Kalita, R., Sudhakar, S.: Hyperspectral remote sensing classifications: a perspective survey. Trans. GIS **20**(4), 463–490 (2015)
36. Clark, R.N., Roush, T.L.: Reflectance spectroscopy: quantitative analysis techniques for remote sensing applications. J. Geophys. Res. **89**(7), 6329–6340 (1984)
37. Cloutis, E.A.: Hyperspectral geological remote sensing: evaluation of analytical techniques. Int. J. Remote Sens. **17**(12), 2215–2242 (1996)
38. Craig, M.D.: Minimum-volume transforms for remotely sensed data. IEEE Trans. Geosci. Remote Sens. **32**, 542–552 (1994)
39. Fauvel, M.: Spectral and Spatial Methods for the Classification of Urban Remote Sensing Data. Ph.D. thesis, Grenoble Institute of Technology (2007)
40. Fauvel, M., Chanussot, J., Benediktsson, J.A.: A spatial-spectral kernel-based approach for the classification of remote-sensing images. Pattern Recogn. **45**(1), 381–392 (2012)
41. Fauvel, M., Tarabalka, Y., Benediktsson, J.A., Chanussot, J., Tilton, J.C.: Advances in spectral-spatial classification of hyperspectral images. Proc. IEEE **101**(3), 652–675 (2013)
42. Geman, S., Geman, D.: Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. IEEE Trans. Pattern Anal. Mach. Intell. **6**(6), 721–741 (1984)
43. Ghiyamat, A., Shafri, H.Z.M.: A review on hyperspectral remote sensing for homogeneous and heterogeneous forest biodiversity assessment. Int. J. Remote Sens. **31**(7), 1837–1856 (2010)
44. Gillespie, A.R., Smith, M.O., Adams, J.B., Willis, S.C., Fisher, A.F., Sabol, D.E.: Interpretation of residual images: spectral mixture analysis of AVIRIS images, Owens Valley, California. In: Green, R.O. (ed.) Proceedings of the 2nd AVIRIS Workshop, vol. 90–54, pp. 243–270 (1990)
45. Gökhan Akçay, H., Aksoy, S.: Building detection using directional spatial constraints. In: IEEE IGARSS, pp. 1932–1935 (2010)
46. Gonzalez, R.C., Woods, R.E.: Digital Image Processing, 2nd edn. Prentice Hall, Englewood Cliffs (2002)
47. Green, R.O., Eastwood, M.L., Sarture, C.M., Chrien, T.G., Aronsson, M., Chippendale, B.J., Faust, J.A., Pavri, B.E., Chovit, C.J., Solis, M., et al.: Imaging spectroscopy and the airborne visible/infrared imaging spectrometer (AVIRIS). Remote Sens. Environ. **65**(3), 227–248 (1998)
48. Gualtieri, J.A., Cromp, R.F.: Support vector machines for hyperspectral remote sensing classification. Proc. SPIE **3584**, 221–232 (1998)
49. Guilfoyle, K.J., Althouse, M.L., Chang, C.-I.: A quantitative and comparative analysis of linear and nonlinear spectral mixture models using radial basis function neural networks. IEEE Trans. Geosci. Remote Sens. **39**, 2314–2318 (2001)
50. Ham, J., Chen, Y., Crawford, M.M., Ghosh, J.: Investigation of the random forest framework for classification of hyperspectral data. IEEE Trans. Geosci. Remote Sens. **43**(3), 492–501 (2005)
51. Hapke, B.: Bidirection reflectance spectroscopy. I. theory. J. Geophys. Res. **86**, 3039–3054 (1981)

52. Hapke, B.: Theory of Reflectance and Emittance Spectroscopy. Cambridge University Press, Cambridge (1993)
53. Harsanyi, J., Farrand, W., Chang, C.-I.: Determining the number and identity of spectral endmembers: an integrated approach using Neyman–Pearson eigenthresholding and iterative constrained RMS error minimization. Proc. Them. Conf. Geol. Remote Sens. **1**, 1–10 (1993)
54. Harsanyi, J.C., Chang, C.-I.: Hyperspectral image classification and dimensionality reduction: an orthogonal subspace projection approach. IEEE Trans. Geosci. Remote Sens. **32**(4), 779–785 (1994)
55. Heinz, D., Chang, C.-I.: Fully constrained least squares linear mixture analysis for material quantification in hyperspectral imagery. IEEE Trans. Geosci. Remote Sens. **39**, 529–545 (2001)
56. Heinz, D.C., Chang, C.-I., Althouse, M.L.G.: Fully constrained least squares-based linear unmixing. IEEE IGARSS **1**, 1401–1403 (1999)
57. Hendrix, E.M.T., Garcia, I., Plaza, J., Martin, G., Plaza, A.: A new minimum volume enclosing algorithm for endmember identification and abundance estimation in hyperspectral data. IEEE Trans. Geosci. Remote Sens. **50**(2), 2744–2757 (2012)
58. Hu, Y.H., Lee, H.B., Scarpace, F.L.: Optimal linear spectral unmixing. IEEE Trans. Geosci. Remote Sens. **37**, 639–644 (1999)
59. Hughes, G.: On the mean accuracy of statistical pattern recognizers. IEEE Trans. Inf. Theory **14**(1), 55–63 (1968)
60. Iordache, M.D., Bioucas-Dias, J., Plaza, A.: Sparse unmixing of hyperspectral data. IEEE Trans. Geosci. Remote Sens. **49**(6), 2014–2039 (2011)
61. Jackson, Q., Landgrebe, D.: Adaptive bayesian contextual classification based on Markov random fields. IEEE Trans. Geosci. Remote Sens. **40**(11), 2454–2463 (2002)
62. Jimenez, L.O., Landgrebe, D.A.: Supervised classification in high-dimensional space: geometrical, statistical, and asymptotical properties of multivariate data. IEEE Trans. Syst. Man Cybern. **28**(1), 39–54 (1998)
63. Keshava, N., Kerekes, J., Manolakis, D., Shaw, G.: An algorithm taxonomy for hyperspectral unmixing. In: Proceedings of the SPIE AeroSense Conference on Algorithms for Multispectral and Hyperspectral Imagery VI, vol. 4049, pp. 42–63 (2000)
64. Keshava, N., Mustard, J.F.: Spectral unmixing. IEEE Signal Process. Mag. **19**(1), 44–57 (2002)
65. Kettig, R.L., Landgrebe, D.A.: Classification of multispectral image data by extraction and classification of homogeneous objects. IEEE Trans. Geosci. Electron. **14**(1), 19–26 (1976)
66. Landgrebe, D.: Hyperspectral image data analysis. IEEE Signal Proces. Mag. **1053–5888**, 17–28 (2002)
67. Landgrebe, D.A.: Signal Theory Methods in Multispectral Remote Sensing. Wiley, New York (2003)
68. Landgrebe, D.A.: Multispectral land sensing: where from, where to? IEEE Trans. Geosci. Remote Sens. **43**(3), 414–421 (2005)
69. Lassalle, P., Inglada, J., Michel, J., Grizonnet, M., Malik, J.: A scalable tile-based framework for region-merging segmentation. IEEE Trans. Geosci. Remote Sens. **53**(10), 5473–5485 (2015)
70. Lee, J.B., Woodyatt, S., Berman, M.: Enhancement of high spectral resolution remote-sensing data by noise-adjusted principal components transform. IEEE Trans. Geosci. Remote Sens. **28**(3), 295–304 (1990)
71. Li, J., Bioucas-Dias, J.: Minimum volume simplex analysis: a fast algorithm to unmix hyperspectral data. IEEE IGARSS **3**, 250–253 (2008)
72. Li, J., Bioucas-Dias, J.M., Plaza, A.: Spectral-spatial hyperspectral image segmentation using subspace multinomial logistic regression and markov random fields. IEEE Trans. Geosci. Remote Sens. **50**(3), 809–832 (2013)
73. Li, J., Bioucas-Dias, M., Plaza, A.: Collaborative nonnegative matrix factorization for remotely sensed hyperspectral unmixing. In: IEEE IGARSS (2012)

74. Liao, W., Dalla Mura, M., Chanussot, J., Bellens, R., Philips, W.: Morphological attribute profiles with partial reconstruction. IEEE Trans. Geosci. Remote Sens. (2015)
75. Ma, X., Jie, G., Hongyu, W.: Hyperspectral image classification via contextual deep learning. EURASIP J. Image Video Process. **2015**(1), 1–12 (2015)
76. Maggiori, E., Tarabalka, Y., Charpiat, G.: Improved partition trees for multi-class segmentation of remote sensing images. In: IEEE IGARSS, pp. 1016–1019. IEEE (2015)
77. Maggiori, E., Tarabalka, Y., Charpiat, G.: Optimizing partition trees for multi-object segmentation with shape prior. In: 26th British Machine Vision Conference (2015)
78. Mazer, A.S., Martin, M.: Image processing software for imaging spectrometry data analysis. Remote Sens. Environ. **24**(1), 201–210 (1988)
79. Melgani, F., Bruzzone, L.: Classification of hyperspectral remote sensing images with support vector machines. IEEE Trans. Geosci. Remote Sens. **42**(8), 1778–1790 (2004)
80. Merényi, E.: Intelligent understanding of hyperspectral images through self-organizing neural maps. In: Proceedings of the 2nd International Conference on Cybernetics and Information Technologies, Systems and Applications (CITSA 2005), Orlando, FL, USA, pp. 30–35 (2005)
81. Miao, L., Qi, H.: Endmember extraction from highly mixed data using minimum volume constrained nonnegative matrix factorization. IEEE Trans. Geosci. Remote Sens. **45**(3), 765–777 (2007)
82. Moser, G., Serpico, S.B.: Combining support vector machines and Markov random fields in an integrated framework for contextual image classification. IEEE Trans. Geosci. Remote Sens. **51**(5), 2734–2752 (2013)
83. Moser, G., Serpico, S.B., Benediktsson, J.A.: Land-cover mapping by Markov modeling of spatial-contextual information in very-high-resolution remote sensing images. Proc. IEEE **101**(3), 631–651 (2013)
84. Nascimento, J.M.P., Bioucas-Dias, J.M.: Vertex component analysis: a fast algorithm to unmix hyperspectral data. IEEE Trans. Geosci. Remote Sens. **43**(4), 898–910 (2005)
85. Parente, M., Plaza, A.: Survey of geometric and statistical unmixing algorithms for hyperspectral images. In: IEEE WHISPERS, pp. 1–4 (2010)
86. Patra, S., Bruzzone, L.: A batch-mode active learning technique based on multiple uncertainty for svm classifier. IEEE Geosci. Remote Sens. Lett. **9**(3), 497–501 (2012)
87. Petrou, M., Foschi, P.G.: Confidence in linear spectral unmixing of single pixels. IEEE Trans. Geosci. Remote Sens. **37**, 624–626 (1999)
88. Plaza, A., Benediktsson, J.A., Boardman, J., Brazile, J., Bruzzone, L., Camps-Valls, G., Chanussot, J., Fauvel, M., Gamba, P., Gualtieri, J.A., Marconcini, M., Tilton, J.C., Trianni, G.: Recent advances in techniques for hyperspectral image processing. Remote Sens. Environ. **113**(Supplement 1), S110–S122 (2009)
89. Plaza, A., Chang, C.-I.: Impact of initialization on design of endmember extraction algorithms. IEEE Trans. Geosci. Remote Sens. **44**(11), 3397–3407 (2006)
90. Plaza, A., Martin, G., Plaza, J., Zortea, M., Sanchez, S.: Recent developments in spectral unmixing and endmember extraction. In: Prasad, S., Bruce, L.M., Chanussot, J. (eds.) Optical Remote Sensing, chap. 12, pp. 235–267. Springer, Berlin (2011)
91. Plaza, A., Martinez, P., Perez, R., Plaza, J.: Spatial/spectral endmember extraction by multidimensional morphological operations. IEEE Trans. Geosci. Remote Sens. **40**(9), 2025–2041 (2002)
92. Plaza, A., Martinez, P., Perez, R., Plaza, J.: A quantitative and comparative analysis of endmember extraction algorithms from hyperspectral data. IEEE Trans. Geosci. Remote Sens. **42**(3), 650–663 (2004)
93. Plaza, A., Martinez, P., Plaza, J., Perez, R.: Dimensionality reduction and classification of hyperspectral image data using sequences of extended morphological transformations. IEEE Trans. Geosci. Remote Sens. **43**(3), 466–479 (2005)
94. Plaza, J., Hendrix, E.M.T., Garcia, I., Martin, G., Plaza, A.: On endmember identification in hyperspectral images without pure pixels: a comparison of algorithms. J. Math. Imaging Vision **42**(2–3), 163–175 (2012)

95. Plaza, J., Plaza, A., Perez, R., Martinez, P.: On the use of small training sets for neural network-based characterization of mixed pixels in remotely sensed hyperspectral images. Pattern Recogn. **42**, 3032–3045 (2009)
96. Qian, Y., Jia, S., Zhou, J., Robles-Kelly, A.: Hyperspectral unmixing via l1/2 sparsity-constrained nonnegative matrix factorization. IEEE Trans. Geosci. Remote Sens. **49**(11), 4282–4297 (2011)
97. Ratle, F., Camps-Valls, G., Weston, J.: Semisupervised neural networks for efficient hyperspectral image classification. IEEE Trans. Geosci. Remote Sens. **48**(5), 2271–2282 (2010)
98. Ren, H., Chang, C.-I.: Automatic spectral target recognition in hyperspectral imagery. IEEE Trans. Aerosp. Electron. Syst. **39**(4), 1232–1249 (2003)
99. Richards, J.A.: Analysis of remotely sensed data: the formative decades and the future. IEEE Trans. Geos. Remote Sens. **43**(3), 422–432 (2005)
100. Rogge, D.M., Rivard, B., Zhang, J., Sanchez, A., Harris, J., Feng, J.: Integration of spatial-spectral information for the improved extraction of endmembers. Remote Sens. Environ. **110**(3), 287–303 (2007)
101. Rubner, Y., Tomasi, C., Guibas, L.J.: A metric for distributions with applications to image databases. In: ICCV, pp. 59–66 (1998)
102. Salembier, P., Foucher, S., López-Martínez, C.: Low-level processing of PolSAR images with binary partition trees. In: IEEE IGARSS (2014)
103. Salembier, P., Garrido, L.: Binary partition tree as an efficient representation for image processing, segmentation, and information retrieval. IEEE Trans. Image Process. **9**(4), 561–576 (2000)
104. Scholkopf, B., Smola, A.J.: Learning with Kernels. MIT Press, Cambridge (2002)
105. Settle, J.J.: On the relationship between spectral unmixing and subspace projection. IEEE Trans. Geosci. Remote Sens. **34**, 1045–1046 (1996)
106. Settle, J.J., Drake, N.A.: Linear mixing and the estimation of ground cover proportions. Int. J. Remote Sens. **14**, 1159–1177 (1993)
107. Shaw, G., Burke, H.: Spectral imaging for remote sensing. Linc. Lab. J. **14**(1), 3–28 (2003)
108. Singer, R.B., McCord, T.B.: Mars: large scale mixing of bright and dark surface materials and implications for analysis of spectral reflectance. In: Proceedings of the Lunar and Planetary Science and Conference, pp. 1835–1848 (1979)
109. Soille, P.: Morphological Image Analysis, 2nd edn. Springer, Heidelberg (2003)
110. Soille, P.: Beyond self-duality in morphological image analysis. Image Vis. Comput. **23**(2), 249–257 (2005)
111. Subramanian, S., Gat, N., Sheffield, M., Barhen, J., Toomarian, N.: Methodology for hyperspectral image classification using novel neural network. Proc. SPIE **3071**, 128–137 (1997)
112. Tarabalka, Y., Benediktsson, J.A., Chanussot, J.: Spectral-spatial classification of hyperspectral imagery based on partitional clustering techniques. IEEE Trans. Geos. Remote Sens. **47**(9), 2973–2987 (2009)
113. Tarabalka, Y., Benediktsson, J.A., Chanussot, J., Angulo, J., Fauvel, M.: Classification of hyperspectral data using support vector machines and adaptive neighborhoods. In: Proceedings of the 6th EARSeL SIG is Workshop (2009)
114. Tarabalka, Y., Benediktsson, J.A., Chanussot, J., Tilton, J.C.: Multiple spectral-spatial classification approach for hyperspectral data. IEEE Trans. Geosci. Remote Sens. **48**(11), 4122–4132 (2010)
115. Tarabalka, Y., Chanussot, J., Benediktsson, J.A.: Segmentation and classification of hyperspectral images using minimum spanning forest grown from automatically selected markers. IEEE Trans. Syst. Man Cybern. Part B **40**(5), 1267–1279 (2010)
116. Tarabalka, Y., Chanussot, J., Benediktsson, J.A.: Segmentation and classification of hyperspectral images using watershed transformation. Pattern Recogn. **43**(7), 2367–2379 (2010)
117. Tarabalka, Y., Fauvel, M., Chanussot, J., Benediktsson, J.A.: SVM- and MRF-based method for accurate classification of hyperspectral images. IEEE GRSL **7**(4), 736–740 (2010)
118. Tarabalka, Y., Rana, A.: Graph-cut-based model for spectral-spatial classification of hyperspectral images. In: IEEE IGARSS, pp. 3418–3421. IEEE (2014)

119. Tuia, D., Camps-Valls, G.: Urban image classification with semisupervised multiscale cluster kernels. IEEE J. Sel. Topics Appl. Earth Obs. Remote Sens. **4**(1), 65–74 (2011)
120. Tuia, D., Volpi, M., Copa, L., Kanevski, M., Muñoz-Marí, J.: A survey of active learning algorithms for supervised remote sensing image classification. IEEE J. Sel. Topics Signal Process. **5**(3), 606–617 (2011)
121. Vaiphasa, C.: Innovative genetic algorithm for hyperspectral image classification. In: Proceedings of the International Conference on Map Asia, vol. 20 (2003)
122. Valero, S., Salembier, P., Chanussot, J.: Hyperspectral image representation and processing with binary partition trees. IEEE Trans. Image Process. **22**(4), 1430–1443 (2013)
123. Valero, S., Salembier, P., Chanussot, J.: Object recognition in urban hyperspectral images using binary partition tree representation. In: IEEE IGARSS (2013)
124. Vapnik, V.: Statistical Learning Theory. Wiley, New York (1998)
125. Wan, L., Tang, K., Li, M., Zhong, Y., Qin, A.K.: Collaborative active and semisupervised learning for hyperspectral remote sensing image classification. IEEE Trans. Geosci. Remote Sens. **53**(5), 2384–2396 (2015)
126. Wang, J., Chang, C.-I.: Applications of independent component analysis in endmember extraction and abundance quantification for hyperspectral imagery. IEEE Trans. Geosci. Remote Sens. **44**(9), 2601–2616 (2006)
127. Wax, M., Kailath, T.: Detection of signals by information theoretic criteria. IEEE Trans. Acoust. Speech Signal Process. **33**(2), 387–392 (1985)
128. Wu, T.-F., Lin, C.-J., Weng, R.C.: Probability estimates for multi-class classification by pairwise coupling. J. Mach. Learn. Res. **5**, 975–1005 (2004)
129. Yang, H., Meer, F.V.D., Bakker, W., Tan, Z.J.: A back-propagation neural network for mineralogical mapping from AVIRIS data. Int. J. Remote Sens. **20**(1), 97–110 (1999)
130. Yuhas, R.H., Goetz, A.F.H., Boardman, J.W.: Discrimination among semi-arid landscape endmembres using the spectral angle mapper (SAM) algorithm. Proc. Ann. JPL Airborne Geosci. Workshop **1**, 147–149 (1992)
131. Zare, A., Gader, P.: Sparsity promoting iterated constrained endmember detection in hyperspectral imagery. IEEE Geosci. Remote Sens. Lett. **4**(3), 446–450 (2007)
132. Zare, A., Gader, P.: Hyperspectral band selection and endmember detection using sparsity promoting priors. IEEE Geosci. Remote Sens. Lett. **5**(2), 256–260 (2008)
133. Zymnis, A., Kim, S.-J., Skaf, J., Parente, M., Boyd, S.: Hyperspectral image unmixing via alternating projected subgradients. In: 41st Asilomar Conference on Signals, Systems, and Computers (2007)

Chapter 3
Very High Spatial Resolution Optical Imagery: Tree-Based Methods and Multi-temporal Models for Mining and Analysis

Fabio Pacifici, Georgios K. Ouzounis, Lionel Gueguen, Giovanni Marchisio and William J. Emery

Abstract This chapter presents a selection of image representation methods for very high spatial resolution optical data acquired by agile satellite platforms. Each method aims at specific properties of the image information content and can be tailored to address unique features in spatial, temporal, and angular acquisitions. Techniques for the identification and characterization of surface structures and objects often employ spatial and spectral features best represented in panchromatic and multi-spectral images, respectively. In both cases, the vastness of the data space can only be addressed effectively by means of some data representation structure that organizes the image information content in meaningful ways. The latter suggest that a globally optimal representation of the object(s) of interest can be obtained through interactions with a scale space as opposed to single-scale information layer. Two examples, the Max-Tree and Alpha-Tree algorithms, are discussed in the context of interactive big data information mining. Optical and structural properties of the surface materials can be exploited by analyzing the tempo-angular domain by means of anisotropic decompositions, which rely on the availability of surface reflectance data and dense angular sampling. Bidirectional reflectance distribution functions of various materials are discussed in detail, showing that the temporal information should always be coupled to the corresponding angular component to make the best use of the available imagery.

Keywords Graph-theoretic methods · RPV decomposition · Tempo-angular anisotropic models · Tree-based methods

F. Pacifici (✉) · G.K. Ouzounis · L. Gueguen · G. Marchisio
DigitalGlobe Inc., Westminster, CO, USA

W.J. Emery
University of Colorado, Boulder, CO, USA

G. Moser and J. Zerubia (eds.), *Mathematical Models for Remote Sensing Image Processing*, Signals and Communication Technology,
https://doi.org/10.1007/978-3-319-66330-2_3

3.1 Introduction

During the last two decades, significant progress has been made in developing and launching satellites with instruments, in both the optical/infrared and microwave regions of the spectra, well suited for Earth observation with an increasingly finer spatial, spectral, and temporal resolution. As an example, WorldView-3 is the first multi-payload, super-spectral, very high spatial resolution commercial satellite. Operating at an altitude of 617 km, WorldView-3 provides 31 cm panchromatic, 1.24 m visible and near-infrared, and 3.7 m short-wave infrared imagery (up to 680,000 km^2 per day), with an average revisit time of less than 1 day.

Sensors with sub-metric resolution allow the detection of small-scale objects, such as elements of residential housing, commercial buildings, transportation systems, and utilities. Spectral capabilities provide additional discriminative features for classes that are spatially similar, due to their higher spectral resolution. The temporal component, integrated with the spectral and spatial dimensions, provides essential information, for example, on vegetation dynamics. Finally, newer classes of satellites have high-performance camera control system capable of rapid re-targeting, allowing the collection of dozens of images over a single target, each with a unique angular perspective, opening a unique approach to multi-temporal/multi-angular imaging. However, the current offering of imagery does not match the customer real need. Users in all domains require information or information-related services that are focused, concise, reliable, low cost, timely, and which are provided in forms and formats compatible with the user's own activities. The information extraction process is generally too complex, too expensive, and too dependent on user conjecture to be applied systematically over an adequate number of scenes. Therefore, there is the need to develop fully automated techniques.

While spectral information allows the discrimination of many materials, very high spatial resolution data enable the description of ground structures that are directly linked to image segments. This mapping between image segments and ground structures can be further automatized with the use of image representation structures and machine learning methods. The synergy of the two domains is challenged to address the well-known problem of single-scale segmentation in capturing image structures in their entirety when the latter are inherently multi-scale. Moreover, the computational complexity of the methods involved in relation to the number of pixels that needs to be processed possesses the need for extremely efficient algorithms. The temporal and angular domains significantly inflate the processing requirements of very high spatial resolution imagery.

This chapter presents key methods and strategies for addressing both challenges. In the object detection domain discussed in Sect. 3.2, image representations are data structures that organize the image information content into hierarchies of nested connected components or connected sets of maximal extent. Two known examples in the field of satellite image analysis are the Max-Tree and Alpha-Tree structures. The Max-Tree data structure is defined for scalar data layers such as panchromatic images and clusters isotone pixels according to some definition of spatial

connectivity. Max-Trees can be utilized for the representation of multi-spectral imagery using adequate dimension reduction or edge transforms. Alpha-Trees capture the plurality of radiometric information and provide highly granular representations of the image content. Progressive clustering of neighboring elements of the spatial domain is subject to one or more spectral or other dissimilarity metrics. Tree nodes in both representations coincide with connected components of the image. Node nesting allows for the efficient computation of spectral and moment-based shape features and further for the definition of efficient processes. The computation of both trees has been optimized for sequential, cluster, and cloud environments, allowing for rapid access to the structured/attributed image information content. Moreover, both trees can be set to coincide with space-partitioning data structures such as the kd-Tree for classification of connected components based on limited training data. More specifically, Sect. 3.2.2 gives an overview of notions of connectivity employed to define connected image regions as the first step to image content organization. The Max-Tree and Alpha-Tree structures encoding hierarchical properties of connected image regions over pre-defined intensity or dissimilarity ranges, respectively, are presented in Sect. 3.2.3. The same section discusses the incremental collection of auxiliary data for the computation of node attributes. A mathematical formulation of spatial sampling facilitated by each tree separately is given in Sect. 3.2.4. Section 3.2.5 discusses the classification based on tree pruning and further presents the kd-Tree structure employed for this purpose. The capabilities of this hybrid protocol utilizing a Max-Tree and a kd-Tree are demonstrated over two data sets in Sect. 3.2.6. Multi-temporal and multi-angular image sets can be analyzed in similar ways using representation forests instead of individual trees. In the case where surface materials are analyzed, image representations become pixel-based transforms, examples of which are discussed in Sect. 3.3. In particular, in Sect. 3.3.2, the differences between non-physical and physical quantities are described in detail, including the theoretical formulation to retrieve at-sensor radiance, at-sensor reflectance, and surface reflectance, while the effects of surface anisotropy and two different angular decomposition models are reviewed in Sect. 3.3.3. Both qualitative and quantitative experimental results are discussed in Sect. 3.3.4 to illustrate the advantages of physical quantities and angular decompositions for the analysis of multi-temporal data sets.

3.2 Interactive Image Information Mining Based on Hierarchical Data Representation Structure Coupling

3.2.1 Introduction

Methods for the analysis of very high spatial resolution (VHsR) remote sensing optical imagery are often confronted with global inconsistencies that emerge from

their dependence on sensor acquisition parameters, increased scene complexity, high spatial resolution, and vast region coverage. In response to this, data organization schemes and machine learning have been introduced that led to a new paradigm of pixel- or object-based classification [1–3]. A key element in the success of this paradigm is the robustness and efficiency of the semantic bridge built between automatically derived image features and human knowledge [4]. The latter is often represented by limited subsets of class examples. Giving emphasis on improving the classification accuracy while maintaining small training sets [5], a human–computer interaction model has been proposed in which the user assigns manually class labels on samples suggested by the computer [6, 7]. This makes it adaptive to the user's interests and allows for fine-tuning. Examples of this approach utilize complex modeling schemes that generally scale up linearly with the number of image elements or objects. Consequently, the application of the derived model to the full extent of the image data often proves to be a bottleneck when analyzing massive image tiles. Recomputing the model for fine-tuning or correcting inaccuracies in such cases tends to be tedious and rather inefficient. Recent works showed that the use of structured data representation by clustering techniques leads to considerable speedups in classification [8–11]. Raw clustering, however, often deviates from an objective semantic representation and further requires a priori knowledge of the number of clusters. The latter can be set automatically [12, 13] subject to some optimization criterion. This shows up an improvement in the accuracy of the resulting semantic layers, yet the data organization strategy remains rigid.

Improvements in the performance of the analysis methods based on this paradigm are driven by two directives: efficient management of the input data and of the respective pool of features that drives the classifiers. The input data can be managed efficiently using image representation schemes. The objective is to reduce the number of entities to be addressed by organizing pixels into homogeneous regions or components based on similarity criteria. Image segmentation has been proposed in [14] though its dependence on a scale parameter for determining the extent of the resulting segments makes it a rigid approach that does not support reconfiguration on demand. By contrast, hierarchical segmentation [15, 16] yields a customized reduction in the number of components from some multi-scale representation of the image content. Hierarchical image representations for multi-scale analysis of the image content and segmentation have been investigated in [17–22] and rely on two popular structures: the *Max-Tree* [23] also known as the *Component-Tree* [24] or the *Alpha-Tree* [16]. In both structures, the node linkage matches the nesting order of connected components from the corresponding image level sets or partitions, respectively. The tree nodes rather than storing component features register a limited set of generic variables [25, 26] from which a rich set of structural, intensity, and connectivity attributes can be computed on the fly. This functionality is valuable for exploring both the spectral and geometrical information content of the image under study. Component features can be exported and organized in a multi-dimensional space directly from the tree structure. Image-to-tree and tree-to-feature space mapping protocols for both types of trees have been presented in [26–29] and promote sets of compact feature spaces

to be used in their entirety as feature vectors either to train decision tree classifiers or to compute supervised segmentations.

Building on these developments, a new modular protocol was proposed in [30] that utilizes a combination of the Max-Tree and kd-Tree structures. The Max-Tree was employed for organizing the input image into meaningful components, and for computing their features. Moreover, the Max-Tree itself was used as an efficient interface between the image space and feature space [31] for managing the collection of positive and negative examples in real time. The kd-Tree is a hierarchical clustering algorithm employed for managing the feature space organization. It offers a structured representation of this space from which a classification is computed directly. This representation was configured with a trade-off parameter that allows the user to select the extent of the clustering granularity. This makes the clusters adapted to the classification problem, and further supports rapid and computationally efficient re-adaptation. The classification is used for selecting the desired components/tree nodes which are then mapped back to the spatial domain through the Max-Tree interface.

As opposed to the classical paradigm of interactive learning that is followed by a time-consuming model application on the spatial domain, the approach in [30] shifts the operational complexity to the structuring of the feature space. Following this stage, interactive classification of massive image data sets can be launched in near real time. Experiments reported in [30] that utilize the proposed protocol on gigapixel-sized images were concluded in 33 min (Max-Tree and kd-Tree construction: 3 and 30 min., respectively), on an eight-core architecture (2.2 GHz Intel Xeon) and 64 Gb RAM. Classification results were produced in approximately 10 s. allowing for an interactive query of the information content. With the hierarchical image and feature space data representation structures stored in memory, scene classification, subject to different criteria, can be reiterated rapidly offering a dynamic view of the massive image information content.

3.2.2 Image Content Organization

Organizing the image information content into meaningful entities requires two preliminaries: the definition of the pixel properties to be explored and the set of connectivity rules [32] based on which pixels can be clustered. Examples of the first preliminary are frequency/reflectance values or vectors and spatial properties. Examples of the second include set or lattice theoretic notions of connectivity [32–37], of hyper-connectivity [33, 35, 38, 39], of attribute space connectivity [40], and others [37, 41]. For reasons of simplicity, the remaining discussion will focus on two notions of set connectivity only, namely the *standard morphological connectivity* expressed through the construct of connectivity classes [32] and a sub-connection of the canonical path-wise connection on graph spaces referred to as *alpha-connectivity* [42].

The objective in both cases is to define connected components [32] or otherwise connected sets of maximal extent within the image definition domain E. The maximality condition ensures that for each connected component $\mathrm{CC} \subseteq E$, there

can be no other superset of it that adheres to the same connectivity rules. Sets of maximal extent assign meaningful boundaries to the image information content and allow for the explicit treatment of the regions they represent. Naturally, any two connected components are strictly disjoint, and under the constraint that the empty set is connected too, the set of all connected components constitutes a unique partition of E.

Connected components can be accessed by a morphological operator (see also Chap. 7 for a review of morphological operators) known as a connectivity opening Γ_x [32, 43, 44]. Given a point $x \in E$, Γ_x returns the connected component containing x or the empty set otherwise. The plurality of pixels constituting each connected component can be used for computing features best describing it or differentiating it from others. Image operators that process connected components are called *connected attribute filters* [45–47] and possess a valuable property; they can either retain the component in question intact or remove it in its entirety depending on one or more of its feature/attribute values.

Connected regions in panchromatic images can be defined using two further entities: the *peak components* and *flat zones* [23, 48]. Let f be a panchromatic image with a definition domain E, i.e., $f : E \rightarrow Z^2$. A peak component $P^h(f)$ at level h and marked by a point x is a connected component of the set of all pixels that have intensity greater than or equal to h:

$$P_x^h(f) = \Gamma_x(\{y \in E \mid f(y) \geq h\}). \tag{3.1}$$

A flat zone at level h addresses a subset of the corresponding peak component. It is a connected component marked by x, of the set of all pixels in E that have an intensity strictly equal to h:

$$F_x^h(f) = \Gamma_x(\{y \in E \mid f(y) = h\}). \tag{3.2}$$

If a peak component defines a single flat zone of the same extent, i.e., $F_x^h(f) = P_x^h(f)$, it is called a regional maximum. A regional maximum at level h has no neighbors of intensity greater than or equal to h.

By contrast to the isotone connected image regions discussed so far, α connectivity offers a different perspective that allows the clustering of non-isotone pixels into connected components: the α-connected components [42].

Let δ be some dissimilarity measure between elements of E such that $\delta(x, x) = 0 \; \forall \; x \in E$. To evaluate δ between any two path-connected but not adjacent elements $xandy \in E$, consider a path π between x, y to be a chain of pairwise adjacent elements given by:

$$\pi(x \rightsquigarrow y) \equiv \langle x = x_0, x_1, \ldots, x_{N-1} = y \rangle, \tag{3.3}$$

in which, N is the number of elements in the path.

If $\Pi \neq \emptyset$ is the set of all possible paths between x and y, and N_π is the number of elements in each path, the minimum dissimilarity measure with respect to some

pre-specified element attribute is the ultra-metric functional $\hat{\delta}$ [16] given by:

$$\hat{\delta}(x, y) = inf_{\pi \in \Pi} \left\{ sup_{i \in [0,\ldots,N_\pi - 1]} \{ \delta(x_i, x_{i+1}) \mid x_i, x_{i+1} \in \pi \} \right\} \tag{3.4}$$

In words, (3.4) states that the dissimilarity measure between any two path-connected elements of E is the infimum among the set of values, each corresponding to the maximal dissimilarity between pairwise adjacent elements along each path [42]. This has been used to define single linkage and α connected components accordingly:

An α connected component $\alpha\text{CC}(x)$ marked by $x \in E$ is given by:

$$\alpha\text{CC}(x) = \{ y \mid \hat{\delta}(x, y) \leq \alpha \}. \tag{3.5}$$

Following the analysis in [16], an α connected component is the union of the marking singleton set $\{x\}$ which is α connected to itself, with the family of all elements y to which x is path-connected to such that the dissimilarity between any two adjacent elements in each path separately is no greater than α. If the dissimilarity between any two adjacent elements x and y is less than or equal to α, the two are directly connected, i.e., there exists an edge between x and y, and thus are members of the same αCC. The case in which $\hat{\delta}(x, y) > \alpha$ does not imply that x and y do not belong to the same αCC but only that there is no direct linkage between them. An example is shown in Fig. 3.1. Connected components that consist exclusively of elements for any two of which $\hat{\delta}(x, y) = 0$ are called *reference connected components* or 0-CCs. In the case in which the element attribute is the intensity and the dissimilarity measure δ is the L_p norm, a reference component marked by x is equivalent to the respective image flat zone containing x, i.e., a maximal connected iso-intensity image region.

Fig. 3.1 *Top row* A six-level test pattern and its threshold decomposition with respect to intensity; *second row* the α-linkage of the test pattern for $\alpha = 0, 1, 2$. *Bottom row* the corresponding Max-Tree and Alpha-Tree structures

The αCCs are equivalence classes on the image definition domain [42]; consequently, the set of αCCs for all $x \in E$ defines a partition of E, i.e., αCCs are both collectively exhaustive and mutually exclusive in E. Moreover:

$$\bigcup_{x \in E} \alpha\mathrm{CC}(x) = E. \tag{3.6}$$

Evidently, greater values of α result in larger αCCs, i.e., produce coarser partitions of E. Consider a point $x \in E$ and a range A of α values. All respective αCCs containing x are ordered with respect to α such that:

$$\alpha_i\mathrm{CC}(x) \subseteq \alpha_j\mathrm{CC}(x), \ \forall\, \alpha_i \preceq \alpha_j, \text{ and } \alpha_i, \alpha_j \in A. \tag{3.7}$$

Figure 3.1 shows an example of a fine to coarse partition evolution for three α levels. The red segments represent linkage relations between pixels for $\alpha = 0, 1, 2$. The nesting order of components with respect to α allows for the definition of a structured representation that is referred to as the Alpha-Tree [16, 27] and is discussed next.

3.2.3 *Hierarchical Image Representation Structures*

The image information content is often organized and managed efficiently using some hierarchical representation structure. An example of the latter is a tree $\mathscr{T}$ which is an acyclic graph having an entry node called the root R. The set of nodes defining $\mathscr{T}$ is denoted by $\mathscr{N} = \{N_i\}$. Following each root-path, the nodes found the furthest away from the root are called the leaves. The set of leaves is denoted by $\mathscr{L} = \{L_i\}$, and $\mathscr{L} \subset \mathscr{N}$. Moreover, given a node N, the set of its children nodes is denoted by $C^N = \{C_i^N\}$.

This section gives a brief overview of two popular tree types for hierarchical image representation known as the Max-Tree and the Alpha-Tree. An example of the latter is shown in Fig. 3.2.

3.2.3.1 The Max-Tree Representation

The Max-Tree [23] also referred to as the Component-Tree [24] is a hierarchical image representation structure that was introduced in the context of attribute filtering [45]. It is a rooted, uni-directed tree with its leaves corresponding to the regional maxima of the image and its root corresponding to the single connected component defining the background. The hierarchical ordering of the nodes encodes the nesting of peak components with respect to the image grayscale range. Reversing the nesting order, i.e., having the brightest component defining the background, and dark components defining the leaves of the tree, leads to the Min-Tree representation.

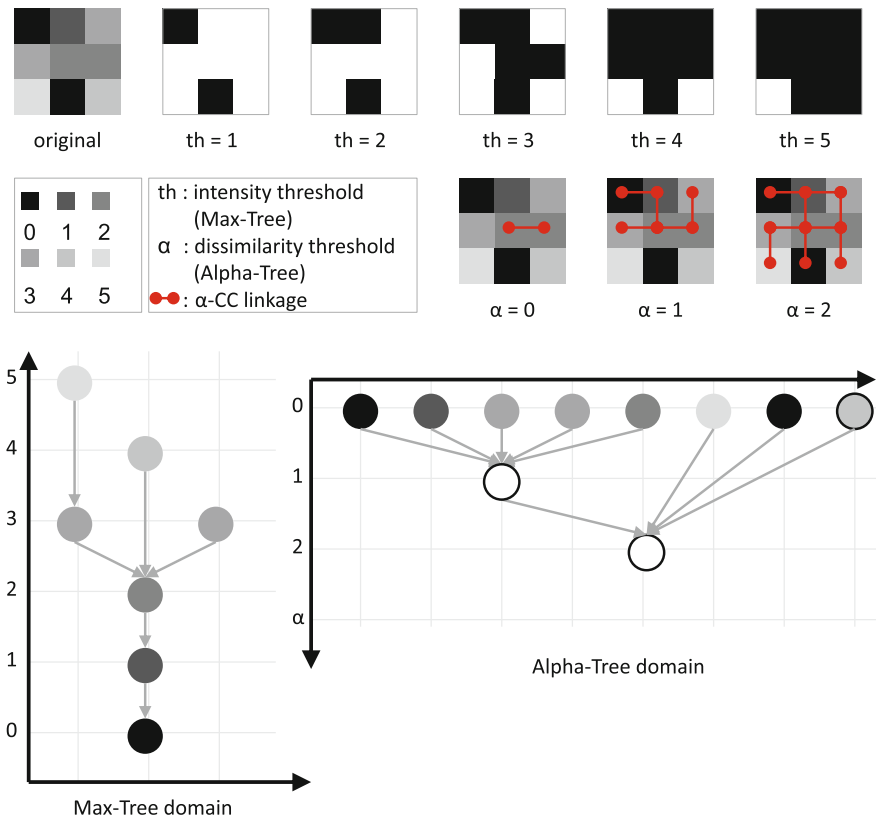


Fig. 3.2 A 3D fish-eye-view projection of an Alpha-Tree domain reduced to a minimum spanning tree. Nodes represented by yellow dots are interconnected in a child-to-parent manner. Points of attraction in this graph space seen here as the tips of various cone-like structures represent parent nodes hosting all children nodes found at the base of each cone (color figure online)

An example of a Max-Tree for a simple six-level test pattern is shown in Fig. 3.1. The top row shows the progressive threshold decomposition of the pattern for five intensity threshold values. Each Max-Tree node N_κ^h associates ($\rightleftarrows$) with a unique peak component and contains all pixels that make up the set of its flat zones. The tree nodes are addressed by their gray level h and the node-at-level index κ:

$$N_\kappa^h \rightleftarrows P_\kappa^h, \tag{3.8}$$

$$N_\kappa^h = \{x \in P_\kappa^h \mid f(x) = h\}. \tag{3.9}$$

The Max-Tree of the test pattern at the top left of Fig. 3.1 is shown at the bottom left diagram. Note that only the peak components containing flat zones are mapped into the tree.

Every node points to its parent which is the first ancestor node below the given level, and the root points to itself. The node structure typically consists of four members: the gray level h, the gray level after node processing h', the parent address, and a pointer to an auxiliary data structure. The latter is variable in size, and it is set upon selecting the types of features to be supported by the tree for further processing. The variables storing the auxiliary data of each node are updated in two separate ways: incrementally after visiting each pixel member of the corresponding set of flat zones and by inheritance from all the children nodes if the reference node is not a regional maximum.

The Max-Tree [23] algorithm runs a three-stage process cycle in which the tree construction is separate from image processing and restitution. The benefit of this modular architecture is that operators may be reiterated without the need for re-computing the tree structure each time.

3.2.3.2 The Alpha-Tree Representation

Consider the set of all αCC partitions, each given by a unique threshold value α on a pre-defined increasing dissimilarity metric. As α increases, partitions turn from fine to coarse following a nesting order. The Alpha-Tree structure [16, 27] is a hierarchical projection of all unique αCCs from this set. Each tree level contains all α-CCs from the corresponding partition at threshold α that appear for the first time. All α'CCs for which $\alpha'\text{CC} = \alpha\text{CC}$ with $\alpha < \alpha'$ are ignored as they only add redundancy. The leaves of the Alpha-Tree correspond to cells of the finest partition, i.e., 0-CCs, and its root to the single node corresponding to the entire image definition domain. An example is shown in Fig. 3.1. The middle row shows the α linkage between elements of the test pattern. For $\alpha = 0$, there exist eight unique 0-CCs. For $\alpha = 1$, there exist four αCCs but only one of them is unique. For $\alpha = 2$, all elements of the definition domain of the test pattern are clustered into a single component defining the root of the Alpha-Tree. Figure 3.2 shows a real Alpha-Tree in which all yellow end-points represent the leaves of the tree. The existing algorithm for computing the Alpha-Tree [16] is designed in the same modular approach as the Max-Tree to allow the separation of the tree construction from the processing stage.

Each node of the tree maps to a unique α connected component from the stack of α-partitions. The relation between α and the tree level h is retrieved from a lookup table that is populated during the tree construction stage by the function $\alpha 2h()$. For $h = \alpha 2h(\alpha)$:

$$N_\kappa^h \rightleftarrows \alpha\text{CC}, \tag{3.10}$$

The subscript κ specifies the node ID at level h.

$$N_{\kappa}^{0} = \{x \in 0\text{-CC}_{\kappa}\}, \text{ and } N_{\kappa}^{h>0} = \bigcup_{x \in E} \alpha' \text{CC}(x) \rightleftarrows N_{\kappa'}^{h'} \mid \alpha' \text{CC}(x) \Subset \alpha \text{CC}(x), \tag{3.11}$$

in which, $\alpha' \text{CC}(x) \Subset \alpha \text{CC}(x)$ means that $\alpha' \text{CC}(x)$ is the largest subset of $\alpha \text{CC}(x)$ containing the point x at $\alpha' < \alpha$. Like with the Max-Tree, each node points to its parent, i.e., the first superset of the given αCC at $\alpha'' > \alpha$, and the root node points to itself. The Alpha-Tree node structure consists of four members: the node offset that is essentially a node identifier, the node parent address, the α level, and a pointer to the auxiliary data structure.

3.2.3.3 Connected Component Attribution

The pool of auxiliary data associated with each Max-Tree or Alpha-Tree node gives access to a rich set of component features that can be computed in advance or upon accessing the node. Further to common features like area, perimeter, and compactness, a wider range of more advanced metrics can be computed like Hu's moments [49], moment of inertia, variance. Emphasizing on the geometrical description of the image information content, an example of the spatial moments is discussed.

Let N_{κ}^{h} be a tree node representing a peak component P_{κ}^{h} in the case of a Max-Tree or an αCC in the case of an Alpha-Tree. Moreover, let B be the number of children nodes each at a level $h_{\beta} > h$ (Max-Tree) or $h_{q} < h$ (Alpha-Tree), for $\beta \in B$. If N_{κ}^{h} is a regional maximum (Max-Tree) or a reference component (Alpha-Tree), it has no children nodes i.e., $B = 0$, and the spatial moments can be computed directly from the set of pixels associated with the respective node:

$$M_{p,q}(N_{\kappa}^{h}) = \sum_{x \in N_{\kappa}^{h}} i_{x}^{p} j_{x}^{q}, \tag{3.12}$$

in which the pair (i_x, j_x) specifies the spatial coordinates of the pixel $x \in E$, and $p \, and \, q$ are positive integers.

In the general case in which $B \geq 0$, the respective feature for a Max-Tree node is computed from the auxiliary data of all pixels associated with the node N_{κ}^{h} and the respective data inherited from all children nodes C:

$$M_{p,q}(N_{\kappa}^{h}) = \sum_{x \in N_{\kappa}^{h}} i_{x}^{p} j_{x}^{q} + \sum_{\beta=0}^{B} M_{p,q}(C_{\kappa'}^{h_{\beta}}), \tag{3.13}$$

in which κ' refers to the node index at level h_{β}.

For an Alpha-Tree node since it consists explicitly of a plurality of pixels all pre-assigned to children nodes, for $B > 0$ Eq. (3.13) can be simplified to:

$$M_{p,q}(N_\kappa^h) = \sum_{\beta=0}^{B} M_{p,q}(C_{\kappa'}^{h_\beta}). \tag{3.14}$$

Inheritance of auxiliary data guarantees that each pixel will be accessed only once for updating the respective members of each node structure, i.e., the process has a linear complexity with respect to the image size. This is an integral step of the Max-Tree/Alpha-Tree construction and does not require additional iterations. Computing node features from the associated pool of auxiliary data reduces to simply substituting the obtained values in each respective formula. Let k be the total number of features supported by a given pool of auxiliary data. Moreover, let $c_t(p,q) \mid t \in [1,\ldots,k]$ be the complexity of computing a feature t for each node. If the total number of nodes is n, then extracting a k-dimensional feature space has a complexity term $\sum_{t=1}^{k} c_t(p,q) \times n$, i.e., $O(kn)$.

In certain applications, further to component features computed directly from the tree representation of the input image f, it is necessary to incorporate features computed from separate sources or processes. Examples are elevation and color properties following image pan-sharpening (see also Chap. 6).

Let $g : E \to R^2$ be some image representing the external source of information. This information can be injected in a tree node by considering the average or the variance of g from the pixels belonging to it. Denoting with A_g, V_g the accumulators computed incrementally and with a_g the average and v_g the variance of the values of g in the considered node, a recursive formulation similar to (3.13) can be adopted:

$$A_g(N_\kappa^h) = \sum_{x \in N_\kappa^h} g(x) + \sum_{\beta=0}^{B} A_g(C_{\kappa'}^{h_\beta}), \tag{3.15}$$

$$V_g(N_\kappa^h) = \sum_{x \in N_\kappa^h} g(x)^2 + \sum_{\beta=0}^{B} V_g(C_{\kappa'}^{h_\beta}), \tag{3.16}$$

$$a_g(N_\kappa^h) = \frac{A_g(N_\kappa^h)}{M_{0,0}N_\kappa^h}, \tag{3.17}$$

$$v_g(N_\kappa^h) = \frac{V_g(N_\kappa^h)}{M_{0,0}N_\kappa^h} - a_g(N_\kappa^h)^2. \tag{3.18}$$

This process is introduced to allow ingesting other sources of information represented in the form of an image defined on the same grid E as f. It has a linear complexity factor with respect to the number of pixels too.

3.2.4 Spatial Sampling

The introduction of tree structures as intermediate modules between the image space and the feature space [28–31, 50] improves the efficiency of spatial sampling by mapping selected connected components to tree labels and in turn to feature space entries. This is done by selecting regions of interest in the original image and retrieving all the components fully included within them. In this section, a brief mathematical formulation of the image grid to tree mapping is presented for each of the two tree types separately.

3.2.4.1 Spatial Sampling Based on the Max-Tree Structure

Given a gray-level image f, a threshold set of f at level h is a binary image is defined as:

$$T_h(f) = \{x \mid f(x) \geq h\} \tag{3.19}$$

The membership of a pixel $x \in E$ in $T_h(f)$ is given by the characteristic function χ:

$$\chi(T_h(f))(x) = \begin{cases} 1, & \text{if } x \in T_h(f), \quad (3.20) \\ 0, & \text{otherwise.} \quad (3.21) \end{cases}$$

For each threshold set $T_h(f)$ with $h \in h_{min} + 1, \ldots, h_{max}$, let $\{\mathrm{CC}_\kappa\}$ be the set of all its connected components, each mapped to a unique peak component P_κ^h. The values h_{min} and h_{max} are the intensity extrema of the input image, and κ is an instance of the component index set K^h.

A region of interest (ROI) marked on the input image can be represented by a binary mask $S \subseteq E$. To extract the peak components fully contained within this spatial subset, we employ connectivity openings Γ_x and closings Φ_x. This is for addressing bright components with respect to a dark background, and the inverse, respectively. The operator Γ_x returns the connected component marked by a point $x \in E$ or $\emptyset$ otherwise. The connected operator Γ_x^S given by:

$$\Gamma_x^S(T_h(f)) = \begin{cases} \Gamma_x(T_h(f)), & \text{if } \Gamma_x(T_h(f)) \subseteq S, \quad (3.22) \\ \emptyset, & \text{otherwise.} \quad (3.23) \end{cases}$$

returns the connected component of the threshold set that is marked by x, provided it is a subset of or equal to the mask S, or $\emptyset$ otherwise. Reiterating Γ_x^S for each threshold set, with h in the image intensity range, yields the complete set of peak components fully contained within the spatial subset. To address the intensity of each pixel within S, the "component window" function [31] is defined as the mapping $CW : E \rightarrow R$ given by:

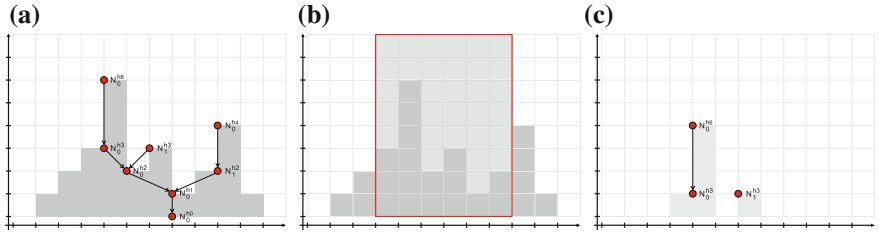


Fig. 3.3 An example of a 1D signal and its Max-Tree in **a**. The y-axis is the intensity axis, and the x-axis is the spatial displacement axis. A ROI in transparent gray in **b** and the contents of the CW and CW^{MT} functions in **c**

$$(CW_S(f))(x) = \sum_{h=h_{min}+1}^{h_{max}} \Big(\sum_{\kappa \in K_h^f} \chi\left(\Gamma_x^S(\mathrm{CC}_\kappa^{T_h(f)})\right)\Big). \tag{3.24}$$

In words, (3.24) states that the intensity of any point $x \in S$, i.e., contained within the component window, can be obtained by initializing it to zero and updating it by adding the value of 1 for each level h at which the connected component $\mathrm{CC}_\kappa^{T_h(f)}$ is a subset of or equal to S and contains the point x.

Let $ID^{MT}()$ be a function that returns the Max-Tree node N_κ^h identifier. To retrieve the identifiers of all the tree nodes that associate with peak components found within S, the function $CW_S^{MT}(f)$ is introduced as:

$$CW_S^{MT}(x) = ID^{MT}(N_\kappa^h) : x \in P_\kappa^h \rightleftarrows N_\kappa^h \text{ and } (CW_S(f))(x) > 0, \tag{3.25}$$

i.e., for each point $x \in E$, $CW_S^{MT}(x)$ returns the index of the node that associates with a peak component containing x such that the component window function in x has a response greater than 0. To compute (3.25), the respective operator is configured with the subtractive filtering rule [26] and requires a double pass through the Max-Tree. The method is described in greater detail in [31].

An example of both functions (3.24) and (3.25) applied over the component window of a simple 1D signal is shown in Fig. 3.3. Image (a) shows six-level signal and the corresponding Max-Tree with the nodes in red. The component window is shown in red frame in image (b). The peak components of the signal that are fully contained within the component window and the corresponding Max-Tree nodes are shown in image (c).

3.2.4.2 Spatial Sampling Based on the Alpha-Tree Structure

The component window function in the case of an Alpha-Tree and for a given point $x \in E$ returns a unique node identifier directly that is given by the function $ID^{AT}()$.

In order to retrieve the maximal node fully contained within the mask S representing the selected ROI, a spatial constraint [16, 51] on the hierarchy of αCCs is required.

Let $\mathbf{P}$ be a logical predicate of αCCs that is defined as:

$$\mathbf{P}(\alpha\mathrm{CC}(x)) = \begin{cases} 1, & \text{if } \alpha\mathrm{CC}(x) \subseteq S, \quad (3.26) \\ 0, & \text{otherwise.} \quad (3.27) \end{cases}$$

The Alpha-Tree component window function is given by:

$$CW_S^{AT}(x) = sup_{h=0}^{h=\alpha 2h(\alpha_{max})} ID^{AT}(N_\kappa^h) : N_\kappa^h \rightleftarrows \alpha\mathrm{CC}(x) \text{ and } \mathbf{P}(\alpha\mathrm{CC}(x)) = 1, \quad (3.28)$$

i.e., for each point $x \in E$, $CW_S^{AT}(x)$ returns the highest Alpha-Tree node index of a node that associates with an αCC containing x (the largest αCC containing x) and for which the spatial predicate is satisfied.

3.2.5 Efficient Classification by Tree Pruning

The representation of an image using either of the Max-Tree or Alpha-Tree structures offers a rich characterization of image components through a limited set of auxiliary data that are associated with each node. In the following, each feature of each component defines an *element*. In application domains like remote sensing image analysis, where the input data are often in the order of gigapixels, the resulting number of objects becomes rather cumbersome to deal with. An example is in classification of elements based on the classical paradigm, i.e., by training a model with a limited subset of data along with the respective class labels and applying it to the full extent of the pool of objects to be classified. In this section, a different paradigm is presented, which profits from a pre-organization of the input elements and leads to a substantial speedup in the classification. It is based on the assumption that any two elements that are close enough with respect to some dissimilarity, i.e., in the same *neighborhood*, are likely to be assigned the same class label. The class label of a single element can be propagated to all other elements in the neighborhood at no further computational cost, thus reducing the overall computational overhead.

3.2.5.1 Classification by Tree Pruning

This approach is known as classification by tree pruning and is inspired from [52]. Assume that we have a set of n k-dimensional elements $\{a_i\}_{i=1}^n$ which are organized in a hierarchical clustering structure, i.e., a tree $\mathscr{T}$. In this analysis, the root is assumed to be at the top of the tree and the leaves at its base. A node N associates with a subset of $\{a_i\}_{i=1}^n$. Given a node N, its children nodes are cover subsets of N that

are not necessary disjoint and for which: $N = \bigcup_i C_i^N$. The root R is the superset containing all elements of $\{a_i\}_{i=1}^n$.

Given a set of training examples $\{\bar{a}_i\}_{i=1}^l$ associated with classes $\{y_i\}_{i=1}^l$, $y_i \in \{1, \ldots, m\}$, the objective is to utilize the tree $\mathscr{T}$ in computing a classification. A naive Bayesian classifier can be built for the elements of each node, where the posterior probability of a random label Y is estimated by assuming a Dirichlet a priori [8, 53], i.e.,

$$p(Y = q \mid N) = \frac{h(y_i = q \mid \bar{a}_i \in N) + 1}{|\{\bar{a}_i \in N\}| + m}. \tag{3.29}$$

The term $h(y_i = q \mid \bar{a}_i \in N)$ is the number of times the class q is represented by the training elements included in the node N. In this way, the elements of a set associated with a node are classified with maximum likelihood label: $\tilde{q} = \arg\max_q p(Y = q \mid N)$. This is a low complexity operation considering that multiple elements are classified in a single step. It is a natural expectation that the elements contained in a same node, being close enough to each other with respect to some dissimilarity, are given the same class label. This modeling of posterior probability has the advantage of being incremental. By maintaining class counters $h(y_i = q \mid \bar{a}_i \in N)$ for each node, each time a new training example becomes available $(\bar{a}_{l+1}, y_{l+1})$, the corresponding element descends along the leaf-paths defined by nodes it belongs to, and updates their counters with its class label y_{l+1}.

To address an element a_i that belongs to a set of nested nodes along a given root-path, all these nodes need to be selected. Consequently, the hierarchical data structure must coincide with a hierarchical data-partition representation. Since the latter is a tree with any given node fully covered by its children, the latter must be mutually disjoint, i.e., $N = \bigcup_i C_i^N$ and $\forall i \neq j, C_i^N \bigcap C_i^N = \emptyset$. In this case, any pruning $\mathscr{P}$ of $\mathscr{T}$ represents a partition of the data set, such that any element a_i belongs to a unique leaf $L_j^{\mathscr{P}}$ of $\mathscr{P}$. Therefore, the element is uniquely classified, by being assigned the class label of the leaf it belongs to.

In the following, a pruning criterion is proposed such that the empirical classification error is minimized while the underlying partition remains coarse. The pruning strategy consists of two stages:

- The tree $\mathscr{T}$ is pruned into $\mathscr{P}_{RD}$ by following a parametrized rate-distortion criterion which defines the finest partition embedded in $\mathscr{P}_{RD}$;
- $\mathscr{P}_{RD}$ is further pruned into $\mathscr{P}$ to minimize the empirical classification error.

In order to minimize the classification error [54], a local criterion minimizing the conditional entropy is suggested:

$$H(Y \mid N) \leq \sum_i \frac{|C_i^N|}{|N|} H(Y \mid C_i^N). \tag{3.30}$$

In the above, $H(Y \mid N) = -\sum_q p(Y = q \mid N) \log p(Y = q \mid N)$. As the conditional entropy reduces, the correct classification rate improves. In a leaf-to-root pass

through the tree, if the local criterion (3.30) is verified, the subtree rooted at node N is pruned. The pass-through terminates for the particular root-path as soon as a node fails the criterion.

A drawback of this strategy, making it computationally inefficient, is that the entire set of leaf nodes $\mathscr{L}$ needs to be considered in the pass through the tree. Moreover, if a leaf contains a single element and this element is part of the training example set, the leaf will be assigned a minimal entropy and will not be pruned. All other leaf nodes having the same parent will not find an example from which classification can be computed. To resolve this shortcoming, an intermediate pruning is performed prior to the classification pruning, such that the intermediate tree leaves are populated sufficiently.

To constrain the extent of the partition granularity, a pruning is performed using a rate-distortion-based criterion [55, 56], justified in the Kolmogorov complexity theory [57]. It is a compression scheme that maps the given elements into an implicit set of feature space vectors, thus reducing the problem of feature selection [58]. A node N consisting of $|N|$ elements requires an average code length of $l(N) = \log \frac{|N|}{|R|}$. Having a metric of the data space, the distortion of the elements belonging to the node is denoted by $d(N)$, and can be approximated by the maximum variance computed between them. If the distortion is small, all elements are very close to each other. As the distance between a node and the root increases, the node distortion gets smaller. The objective of the rate-distortion optimization is to find a balance between the coding length and the average distortion, such that both are minimized. The rate-distortion pruning criterion is given by:

$$\beta\, l(N) + (1-\beta) d(N) \leq \sum_i \frac{|C_i^N|}{|N|} \left(\beta\, l(C_i^N) + (1-\beta) d(C_i^N)\right), \tag{3.31}$$

where β is a trade-off scalar in the range $[0, 1]$. If $\beta = 0$, the parent node is privileged. By contrast, if $\beta = 1$, the node's children are privileged. The rate-distortion criterion is applied on each node in a root-to-leaves pass through the tree. As soon as it is verified, the subtree rooted at N is pruned. The advantage of utilizing this criterion is that non-horizontal cuts of the tree can be computed. Moreover, the trade-off parameter offers the option of selecting the maximal granularity of the pruned tree $\mathscr{P}_{RD}$, before computation of the classification pruning $\mathscr{P}$. Thus, the labeling of all elements is obtained by the naive Bayes classification of the leaves of $\mathscr{P}$. As a remark, the hierarchical clustering structure $\mathscr{T}$ is pruned virtually such that the full structure is maintained. Then, the classification can be re-tuned by considering additional training examples or by modifying the trade-off parameter.

An efficient implementation of a hierarchical clustering algorithm is presented hereafter and is used in the following for performing fast classification by tree pruning.

3.2.5.2 A Hierarchical Clustering Structure: The kd-Tree

Data organization schemes offer considerable advantages regarding the efficiency of accessing elements in vast data spaces. Examples are the hierarchical data representation structures like the hierarchical linkage clustering [27, 59], the vantage-point tree [60], the M-tree [61], or dyadic k-means [62]. An established space-partitioning data structure for organizing elements of a k-dimensional space is the kd-Tree [63]. The tree organizes recursively n data elements of dimension k by applying the following recursion:

- Select the most variant dimension $\hat{k}(N)$ among the k possibilities, considering the elements in the current node N;
- Compute the median $m(N)$ of data in N considering only the most variant dimension;
- Order the elements of the node in two children nodes C_1^N, C_2^N, such that the element under/above the median is in the first/second child, respectively;
- Process the children nodes C_1^N, C_2^N, except if they contain a single element.

This leads to a balanced kd-Tree, in which all leaves are approximately the same distance away from the root. The two children resulting from each parent node contain the same number of elements: $-1 \leq \left|C_1^N\right| - \left|C_2^N\right| \leq 1$. This means that a node at depth $h(N)$ from the root contains at most $2^{\log_2 n - h(N)}$ elements.

An example of a kd-Tree-based space partitioning is shown in Fig. 3.4, for points lying in a two-dimensional space. Initially, the root node is set to contain all the data elements. In the second step, the horizontal dimension is selected for splitting the elements of the root node in two children nodes. The splitting continues recursively in the third and fourth steps, until the leaf nodes are reached, each containing a single

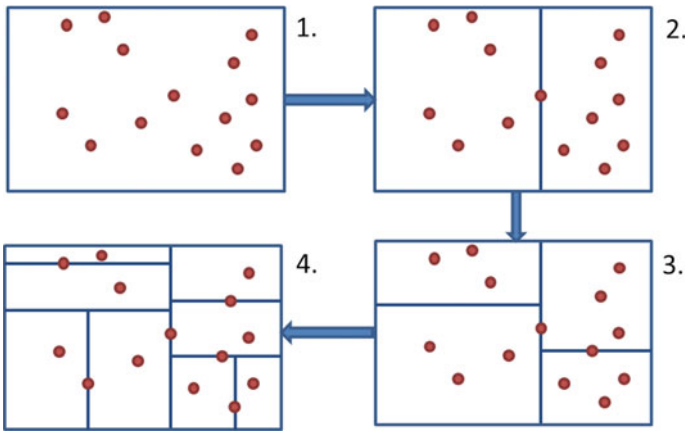


Fig. 3.4 The 4 steps illustrate the progressive construction of the kd-Tree from a data set consisting of two-dimensional elements. It is a recursive process directed from the root to the leaves. In every step, the data space is organized into a finer partition

or a maximum of two data elements. The distortion of a kd-Tree node N is defined as the volume of the hyperrectangle bounding its elements, and it is employed in the computation of the rate-distortion criterion (3.31).

3.2.5.3 Classification Complexity Analysis

In this section, a complexity analysis is given for the proposed classification scheme and a comparison is made against the complexity of the SVM [64] and C4.5 [54] classifiers. It is assumed that the pool of input data contains n samples that have to be classified, each represented by a k-dimensional feature vector. Moreover, a set of m training samples is given.

The data are initially organized into a kd-Tree structure which has a quasi-linear complexity factor of $O(k.n \log_2 n)$. For each new training sample made available, the nearest neighbour is computed and the sample descends from the root to the closest leaf. The classification counter of each node visited along the given path is updated. This operation has a complexity of $O(\log_2 n)$ and is independent of the data dimensionality. In total, the counter update from the set of m training examples has a complexity of $O(m \log_2 n)$. Given a trade-off scalar β, the tree $\mathscr{T}$ is virtually pruned in top-down manner according to (3.31). The complexity of this process is given by the number of nodes that are explored before meeting the pruning criterion. For a well-balanced tree, this is approximated by $O(n^\beta)$. The last stage, i.e., the classification pruning (3.30), is implemented by a further virtual pruning of the intermediate tree in which the class label of its node is propagated to its elements. This has a complexity that is bounded by the number of nodes to be scanned and is given by $O(n^\beta)$.

By contrast to the kd-Tree, the training phase of an SVM (see also Chap. 10) or C4.5 classifiers operated on the same data set have a complexity factor of $O(m^3)$ and $O(m.k.\log_2 m)$, respectively. To classify each data element according to the derived optimal separation, each element is individually evaluated resulting in complexities of $O(n.m.k)$ and $O(n.\log_2 m)$ for the SVM and C4.5 classifiers.

Assuming that $\bar{m}$ training examples are added to the training set, retraining the tree-based classifier requires only $O(\bar{m} \log_2 n)$ operations, while retraining the SVM or the C4.5 classifiers requires to relaunch the process from scratch. These complexity figures are summarized in Table 3.1.

Table 3.1 The algorithmic complexity figures of the full learning process are given for n data elements of dimension k, and for m training samples

	Proposed	SVM	C4.5
Pre-computation	$O(k.n \log_2 n)$	$O(1)$	$O(1)$
Training	$O(m \log_2 n)$	$O(m^3)$	$O(m.k.\log_2 m)$
Classification	$O(n^\beta)$	$O(n.m.k)$	$O(n.\log_2 m)$
Incremental training	$O(\bar{m} \log_2 n)$	$O((m+\bar{m})^3)$	$O((m+\bar{m}).k.\log_2 (m+\bar{m}))$

It is evident from the above that data pre-organization has a strong impact on the classification complexity that reduces sharply compared to the SVM or the C4.5 classifiers. This holds for the case of incremental training too, making the tree-based classification paradigm suitable for an interactive, thus incremental, learning application.

3.2.6 Experiments and Applications

This section demonstrates the proposed classification protocol using a combination of a Max-Tree and kd-Tree structures in two real exercises: the detection of buildings in Port-Au-Prince Haiti, following the Jan. 2010 earthquake, and the detection of refugee camps in Sri Lanka. Both exercises were carried out using an eight-core, 2.2 GHz Intel Xeon machine equipped with 64 Gb of RAM. All algorithms are sequential, i.e., they run on a single core.

The first example makes use of GeoEye-1 image of Port-Au-Prince covering $16 \times 10\,\text{km}^2$ at 0.5 m spatial resolution. It is a panchromatic image acquired in 2010, quantized to 11 bits/pixel and consisting of 33000×20000 pixels. We considered spectral and moment-based features, computed from the complete set of bands available. The image and feature hierarchical representations were computed in 22 m, i.e., 2 m for the Max-Tree and 20 m for the kd-Tree. Running an interactive classification using the tree-based classifier required 3 s. With an adapted visualization strategy, the classification result was mapped back into the image space in 5 s for each iteration of the classifier. In this experiment, two classes were considered, one corresponding to the objects of interest (buildings) and the second to everything else. The classification/detection results obtained after several user–machine interactions are shown in Fig. 3.5.

The second example makes use of a WorldView-1 panchromatic image of refugee camps in Sri Lanka. It is 10000×10000 pixels in size at 0.5 m resolution. In this example, we consider moment-based features alone. The Max-Tree and kd-Tree were computed in approximately in 20 and 200 s., respectively. The classifier for each of the family of targets concluded in 0.3 s, and each result was mapped back into the image domain in 2 s. We classified three types of structures: small tents, big/long tents and buildings and road infrastructure. The results are shown in Fig. 3.6.

To evaluate the accuracy of the tree-based classifier for the two test cases presented, the proposed classification scheme was compared against the C4.5 classifier alone, since the SVM classifier using the libSVM implementation required over an hour to conclude. The C4.5 classifier was based on the J48 weka implementation [65]. It was used for both the training and the classification part, only the latter was optimized by making use of the query capabilities offered by the kd-Tree. That was done to accelerate the C4.5-based query process.

The classification errors were computed by operating each classification scheme on the training data set itself. The results are summarized in Table 3.2. The C4.5 maintained a small lead over the kd-Tree classifier in all cases. Note that since the

Fig. 3.5 Interactive classification of buildings from Port-Au-Prince, Haiti. **a** The *top* image shows the original full scene, from which a subregion (the *blue rectangle*) is selected for training. The *bottom* image shows classification result. **b** The query engine from which positive (*green*) and negative (*red*) ROIs are selected interactively (color figure online)

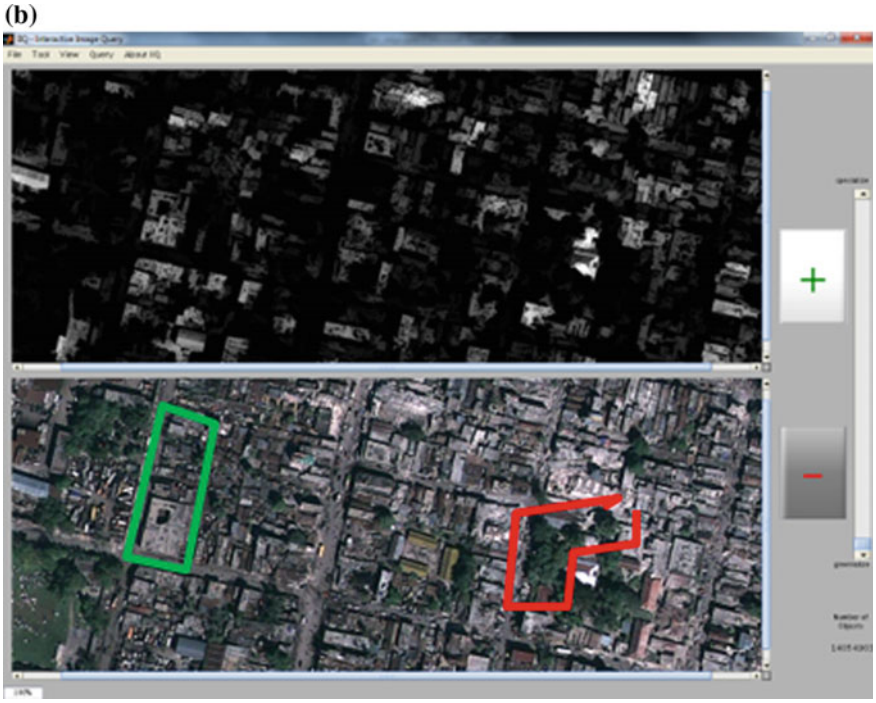


Fig. 3.5 (continued)

training data sets were highly imbalanced, the false-alarm errors were avoided at the expense of more missed detections. As a remark, the reported accuracies were not directly linked to the thematic accuracies; they only give a brief idea of the performance of the proposed algorithm.

A further assessment of the accuracy of the kd-Tree classifier is made by a cross-validation using the feature space projection of the small tents' training data set. The training set is divided in five subsets, such that the classifier training and testing are performed on disjoint sets. The average classification accuracy is evaluated for various values of the trade-off parameter and is shown in Fig. 3.7. It can be seen that an optimal trade-off value exists for which the minimal error probability is obtained.

The error probability against query time is shown in Fig. 3.7. The term P_e decreases sharply for short query times, thus giving an optimal accuracy for a query of no longer than 0.22 s., i.e., 1.5 times the minimum query time. Following this break point, a smooth decrease in the average classification accuracy is observed with respect to the time. Overall, the proposed classifier is seen to operate with an accuracy level close to one of the state-of-the-art algorithms, the C4.5, while being 90 times faster.

A complexity analysis for selected experiments on the full operation cycle of the classification protocol is discussed thoroughly in [30].

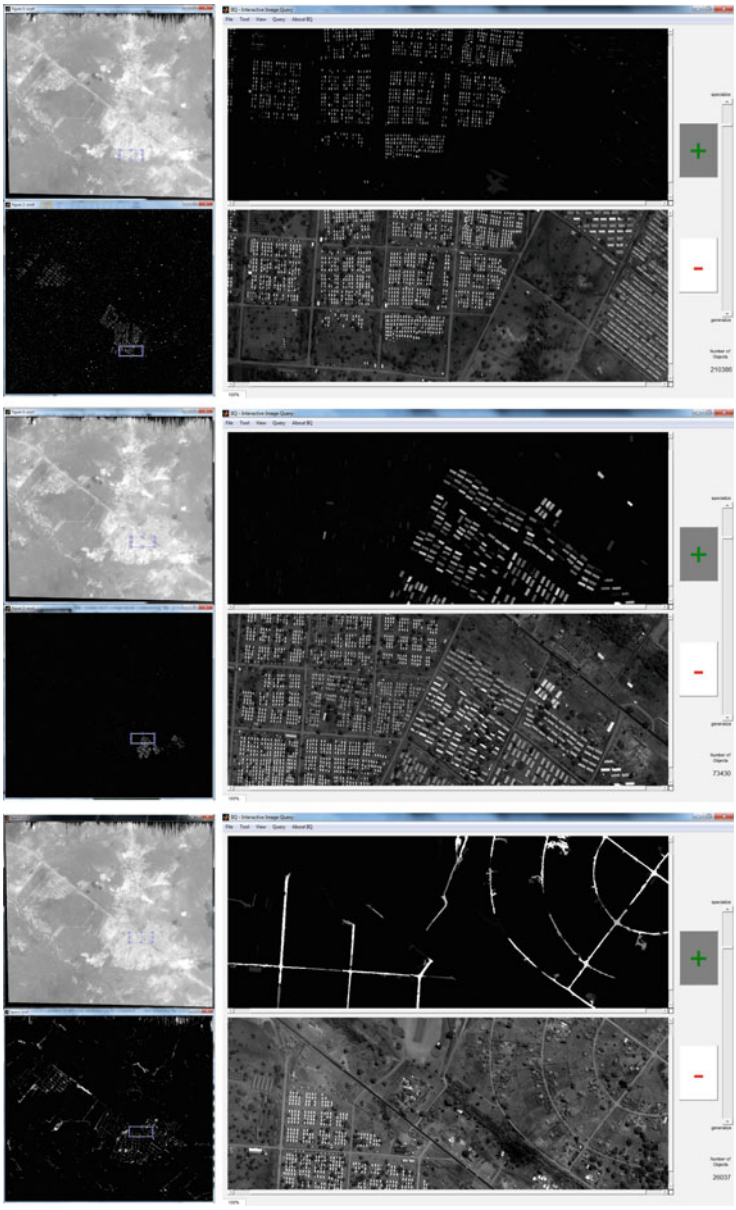


Fig. 3.6 Three interactive classification runs on the image of refugee camps in Sri Lanka. In each of the three sets of images, the *top left* corresponds to the input scene with a region selection marked by a *blue polygon*, the *bottom left* to the final result, and the *right* shows the interface for marking positive and negative samples. The interface shows two views, the *bottom* that corresponds to the selected region from the original image and the results of the classifier as it is fine-tuned by user interaction (color figure online)

Table 3.2 The probability of error P_e, false alarms P_{fa}, and missed detection P_{md} are reported for the two classifiers for each exercise. The query time is the sum of the learning time and the classification of the full input scene consisting of 13,080,502 pixels. The training data set is made of m examples among which Pm are positive

Small tents	P_e	P_{fa}	P_{md}	Query time	m examples	P
kd-Tree classifier	0.036	0.0061	0.37	0.26 s	90513	0.082
C4.5	0.027	0.0088	0.23	15.6 s	90513	0.082
Long tents	P_e	P_{fa}	P_{md}	Query time	m examples	P
kd-Tree classifier	0.16	0.023	0.83	0.52 s	239787	0.16
C4.5	0.159	0.031	0.79	46.4 s	239787	0.16

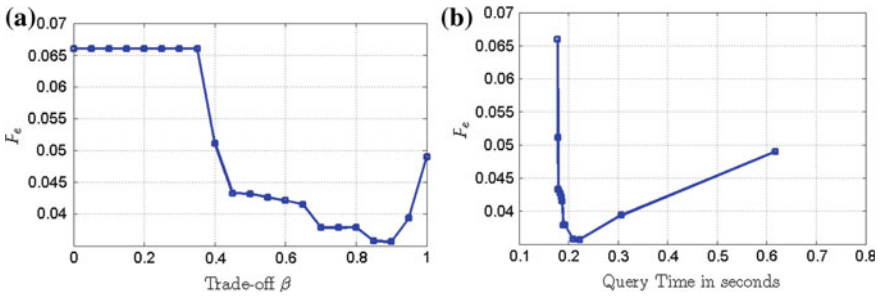


Fig. 3.7 **a** The evolution of the average five fold probability of error P_e with respect to the trade-off parameter β, **b** The evolution of P_e with respect to the query time expressed in seconds

3.3 Mutli-temporal and Multi-angular Optical Image Analysis

3.3.1 Introduction

Commercial availability of optical very high spatial resolution space-borne imagery began more than ten years ago with the launch of IKONOS and QuickBird, which led to an increasing interest in satellite data for mapping and precise location-based service applications. Since then, a large amount of data have been acquired, including images from newer and more complex platforms such as WorldView-1 and WorldView-2, GeoEye-1, and the more recent Pléiades-1A and Pléiades-1B. Currently, the global capacity of the very high spatial resolution imaging satellites is greater than 1.8 billion square kilometers per year (which corresponds to more than 12 times the land surface area of the Earth), and is expected to increase to more than 2.4 billion square kilometers per year (about 16 times the land surface area of the Earth) in the near future. Much of this imagery is collected with a wide range of

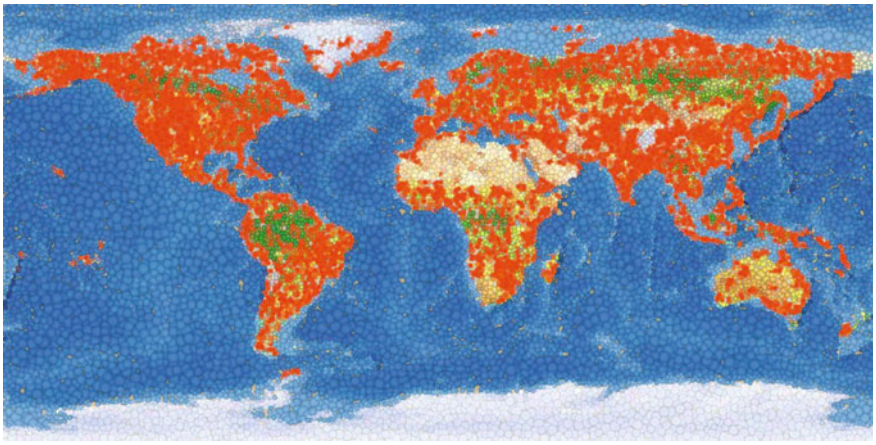


Fig. 3.8 A typical 30-day collection by the QuickBird, WorldView-1, and WorldView-2 constellation

azimuth and elevation angles. Figure 3.8 shows a typical 30-day collection by the QuickBird, WorldView-1, and WorldView-2 constellation.

The availability of sub-meter resolution data regularly acquired over the same geographical region has proved to be effective in developing various monitoring systems, from precision agriculture (including growing and harvesting of crops) to disaster management and search and rescue operations in the case of natural events (earthquakes, hurricanes, floods). In urban areas, multi-temporal data provide information about newly built constructions or demolition of existing structures, road conditions, and urban growth. Further, the WorldView class of satellites has a high-performance camera control system capable of rapid re-targeting, allowing collection within a few seconds of dozens of images over a single target, each with a unique angular perspective. This capability opens a unique approach to multi-temporal imaging whose applications have been discussed in [66].

Despite the large amount of data acquired and the progress of space technology in designing and launching more sophisticated sensors, very little research addresses the advantages and challenges of multi-temporal optical very high spatial resolution data. This has been confirmed by a recent special issue on the analysis of multi-temporal remote sensing data [67], where only one contribution is dealt with sub-meter optical imagery [68].

In the remote sensing literature, two opposite approaches are generally considered to analyze the data sequence in the context of land cover mapping: A set of independent models is developed for each image, or a unique model is generated from the entire set of images at once. The first approach does not guarantee that the temporal information is fully exploited as each classifier is tailored to fit a specific subproblem. On the other hand, a generalized model may suffer from different data distributions in the image sequence, resulting in poor results. In [69], Heas and Datcu propose an

unsupervised method to learn trajectories of dynamic clusters, followed by an interactive learning process. The trajectories in the feature space result in graphs coding spatiotemporal structures contained in the data sequence. The information-bottleneck principle is introduced in [12], which combines model selection with rate-distortion analysis in order to determine the optimal number of clusters. In [70], Petitjean et al. propose an approach to deal with irregularly sampled time-series based on the dynamic-time-warping concept, while various similarity measurements are used to model consecutive image-pairs in [71].

Very recently, there has been a large number of publications in the remote sensing literature on *domain adaptation*, whose goal is to adapt a prediction function from a source domain to a target domain, reducing the effects of *shifts* between different, but related, data sets (such as one gathered from multi-temporal acquisitions). In [72], Bruzzone and Prieto exploit the distribution of a new image to re-estimate the parameters of a maximum likelihood classifier. In [73], a binary hierarchical classifier is used in the target domain to leverage the information extracted from the existing labeled data. In [74], the support vectors of a support vector machine classifier are iteratively adapted to the distribution of a new domain. Active learning methods [75] have also been considered in [76–78] to cope with data set shifts. While most of the domain adaptation methods deal with adjusting the model to the target domain, Tuia et al. propose a method described in [79], where data manifolds are deformed through nonlinear transformations driven by a graph matching procedure aimed at finding correspondences between domains, whereas Leiva-Murillo et al. introduce in [80] the concept of multitask learning by jointly solving a set of prediction problems by sharing information across different tasks. Examples of domain adaptation results will be presented in Chap. 10.

In most remote sensing studies of optical very high spatial resolution imagery, however, the analysis of time-series is limited to the use of pixel digital numbers, ignoring the physical effects of atmospheric, viewing, and illumination changes between image collections. The importance of radiometric calibration (and the need to work with physical quantities) has already been suggested in [81] for the implementation of operational automated remote sensing image understanding systems. As previously discussed in [82], results over the past few years have offered very modest improvements with respect to one obtained from other methods (usually, less than 1–2% in absolute terms) on a limited number of classes, which frequently only include conventional targets, such as man-made structures, vegetation, soil, and water, and number of images (two to three, in general). Often, the various results are not even discussed in terms of their statistical significance leaving the reader wondering if the principles of the proposed technique are repeatable on different data sets or if the improvements were rather obtained by strenuously exercising the data samples to achieve the desired output.

Another limitation of current techniques is that multi-temporal data sets are generally analyzed only considering the temporal domain. Instead, the temporal information should be coupled to the corresponding angular component to make the best use of the available imagery. In fact, the radiometric differences in time-series may often be corrected or accounted for by understanding the physics of the acquisitions.

Table 3.3 Acquisition dates of the time-series with the corresponding day of the year

17-Jul-02 (198)	8-Jan-04 (008)	6-Sep-04 (249)
30-Oct-04 (303)	3-May-05 (123)	14-Jul-05 (195)
27-Jul-05 (208)	4-Sep-05 (247)	7-Oct-05 (280)
25-Nov-05 (329)	5-Feb-06 (036)	18-Mar-06 (077)
14-Jun-06 (165)	30-Jul-06 (211)	5-Nov-07 (309)
8-Nov-07 (312)	18-Mar-08 (077)	23-Mar-08 (082)
22-Aug-08 (234)	31-Jan-09 (031)	12-Aug-09 (224)

The objective of this section is not only to demonstrate that physical quantities are necessary to consistently and efficiently analyze sub-meter optical imagery, but also to bring attention to the research community that the angular information of the acquisitions should not be neglected as unique features can be derived from it.

The data set used is composed of 21 images acquired between 2002 and 2009 by QuickBird (QB) over the city of Denver, Colorado. The time-series covers part of the downtown area and includes single family houses, skyscrapers, apartment complexes, industrial buildings, roads/highways, urban parks, and bodies of water. The acquisition dates are reported in Table 3.3, and the tempo-angular distribution (in zenith and azimuth angles) of the image sequence is shown in Fig. 3.9 along with the Sun position for the specific day of the year (also reported in Table 3.3). All images were acquired within 30 degrees zenith angle and fairly evenly distributed azimuth angles, while the Sun position exhibits the natural declination through the year (being closer to twenty degrees in zenith during summers and to seventy degrees

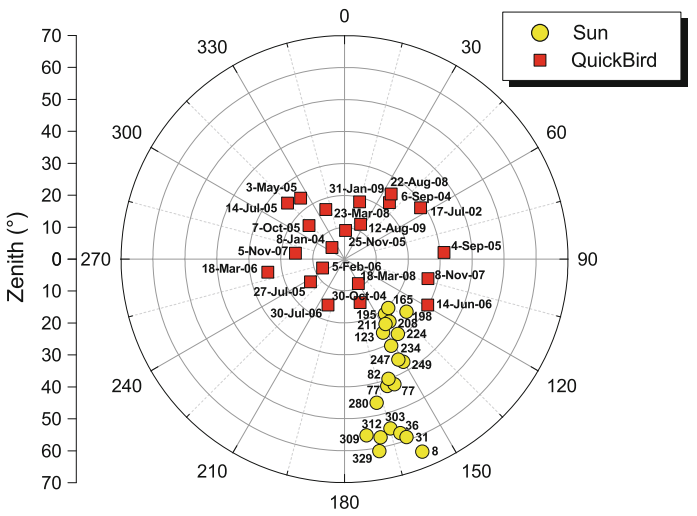


Fig. 3.9 Tempo-angular distribution of the time-series along with the relative Sun position. All images were acquired within 30 degrees zenith angle and fairly evenly distributed azimuth angles. The day of the year of the acquisitions is reported next to the Sun positions

during winters). All 21 images were converted to surface reflectance by AComp, an automatic DigitalGlobe proprietary method designed for very high spatial resolution panchromatic or visible/near-infrared imagery [83] (whose performances are discussed in detail in [84]).

3.3.2 Non-physical and Physical Quantities

The accurate analysis of multi-temporal remote sensing data depends upon the ability to distinguish between relevant and non-relevant changes on the Earth surface through time [85]. At the same time, the capability to detect and quantify these changes depends on consistent sensor measurements, which are generally distorted by changing atmospheric conditions, solar illumination, and satellite viewing geometries. Therefore, it is preferable to convert the raw image counts to physical quantities that are capable of accurately describing the imaged surfaces before analyzing the data from a single scene, between images acquired on different dates, or by different sensors.

With reference to Fig. 3.10, optical remote sensing satellites measure photoelectric signals that are equivalent to the radiance reflected by the Earth surface when illuminated by the Sun and perturbed by the atmosphere, including the effect of gaseous absorption and scattering by molecules and aerosols. These measured signals are not directly accessible to the end users as they are converted and stored as digital numbers (DNs).

The DN values $[counts]$ are proportional to the radiance L $[Wm^{-2}sr^{-1}\mu m^{-1}]$ entering the telescope aperture according to [86]:

$$DN = L \cdot GAIN + OFFSET \tag{3.32}$$

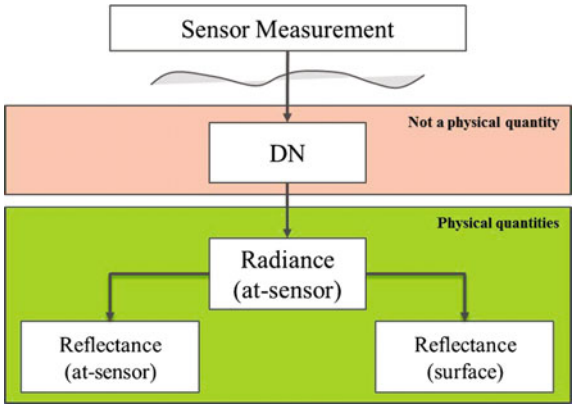


Fig. 3.10 Schematic representation of non-physical and physical quantities. In general, it is preferable to convert the raw image counts to physical quantities that are capable of accurately describing the imaged surfaces before analyzing the data from a single scene, between images acquired on different dates, or by different sensors

where $GAIN$ is the absolute gain $[counts/(Wm^{-2}sr^{-1}\mu m^{-1})]$ and $OFFSET$ is the instrument offset $[counts]$. This formulation assumes that the detectors have a linear response as a function of input radiance. The appropriate gain and offset settings are wavelength-dependent and are operationally selected based on:

- compression levels
- pixel aggregation
- line rate
- bit depth
- time delay integration (TDI), which increases the exposure provided by the basic line rate
- seasonality (the TDI setting for a given image is selected based on the estimated solar elevation angle)

This means that image DN counts are unique not only to the sensor, but also to the very specific operational setting selected for the acquisition. Additionally, an ideal object with a spectral signature that is invariant over time may show significant DN differences in multi-temporal data sets acquired with similar operational settings due to different atmospheric conditions and/or different viewing and solar geometries. In this sense, DN data, which do not represent physical quantities, should not be directly compared to DN imagery from other sensors, nor even between images from the same sensor as collection settings, atmospheric effects, and viewing and illumination geometries may be significantly different.

Calculation of at-sensor, also known as top of atmosphere (TOA), radiance is a necessary step for converting the image into a physically meaningful common scale [87]. Even though end users have access only to the image DN counts, pixel values can easily be converted to at-sensor radiance by inverting Eq. 3.32 and using the $GAIN$ and $OFFSET$ information provided in the image meta-data.

An additional reduction in scene-to-scene variability can be achieved by converting the at-sensor radiance L to TOA reflectance ρ^{TOA} $[unitless]$ using:

$$\rho^{TOA} = \frac{L \cdot d_{ES}^2 \cdot \pi}{E_{sun} \cdot cos(\theta_S)} \tag{3.33}$$

where E_{sun} is the mean exoatmospheric solar irradiance $[Wm^{-2}\mu m^{-1}]$, θ_S is the solar zenith angle $[degrees]$, and d_{ES} is the Earth-Sun distance $[astronomical\ units]$ as derived in the Appendix. There are several benefits for using TOA reflectance with respect to TOA radiance, such as the removal of the cosine effect of different solar zenith angles, the compensation for different values of the exoatmospheric solar irradiance arising from spectral band differences, and the correction for the variation in the Earth-Sun distance between the different acquisitions [87].

To estimate surface reflectance ρ $[unitless]$ from satellite data, TOA radiance needs to be compensated for atmospheric absorption and scattering phenomena, approximating what would be measured by a sensor held just above the Earth surface, without any alterations from the atmosphere [88]. One of the advantages of surface reflectance is the physically based normalization of the image values throughout the

dates regardless of the different atmospheric conditions. Surface reflectance can be derived as:

$$\rho = \frac{(L - L_{up}) \cdot d_{ES}^2 \cdot \pi}{\tau_{up} \cdot (E_{sun} \cdot cos(\theta_S) \cdot \tau_{down} + E_{down})} \tag{3.34}$$

where L_{up} $[Wm^{-2}sr^{-1}\mu m^{-1}]$ is the upward radiance scattered by the atmosphere, τ_{up} $[unitless]$ is the atmospheric transmittance from the ground to the top of the atmosphere, τ_{down} $[unitless]$ is the atmospheric transmittance from the top of the atmosphere to the ground, and E_{down} $[Wm^{-2}\mu m^{-1}]$ is the diffuse irradiance at the surface [89, 90].

The upwelling radiance L_{up} is one of the major components of the atmospheric spectral distortion. Being driven by Rayleigh scattering, its effect has a λ^{-4} dependence on wavelength. In very high spatial resolution satellite imagery, L_{up} is closely related to the minimum radiance measured by the sensor, which generally corresponds to shadowed areas in the image with no direct solar illumination. This means that even shadowed surfaces, regardless of the specific material, can show different values of measured radiance (and, therefore, DN counts) at two different acquisition times due to the dependence of L_{up} to the acquisition conditions (both in terms of Sun-Earth-satellite geometry and/or atmospheric properties).

As a practical example, Fig. 3.11 illustrates spectral signatures of grass for both TOA and surface reflectance (retrieved by the algorithm in [83, 84]) for four urbanized areas acquired by WorldView-2 just after its launch at the end of 2009. The use of WorldView-2 imagery for this example is necessary because of the presence of two relevant bands centered at 427 and 949 nm, which correspond to spectral regions with higher Rayleigh scattering and water vapor absorption, respectively. For each city, two different patches of grass were selected, one from sport facilities, such as stadiums, and another from public parks, ideally corresponding to different levels of vegetative health. It should be noted that each of these acquisitions represents locations across the globe with different climates, seasons, and grass species. An additional hyperspectral signature of grass as provided by publicly available libraries [91] is reported with a red line for comparison.

TOA reflectance signatures are consistently larger than surface reflectance values at shorter wavelengths due to Rayleigh scattering, ranging from about 0.15 to 0.21 at 427 nm (to be compared with both the reference hyperspectral and the retrieved surface reflectance signatures that range from about 0.02 to 0.04). On the other hand, TOA reflectance values at 949 nm are smaller than both the reference hyperspectral and the retrieved surface reflectance signatures due to water vapor absorption. It is important to account for these spectral distortions. For example, the average of the normalized difference vegetation index (NDVI) calculated from TOA reflectance values is about 0.688, to be compared to 0.814 in case of surface reflectance, which corresponds to a difference of more than 15%. Therefore, surface reflectance not only provides consistent quantities across the various bands at different times, geographical locations, and atmospheric conditions, but it also minimizes the spectral

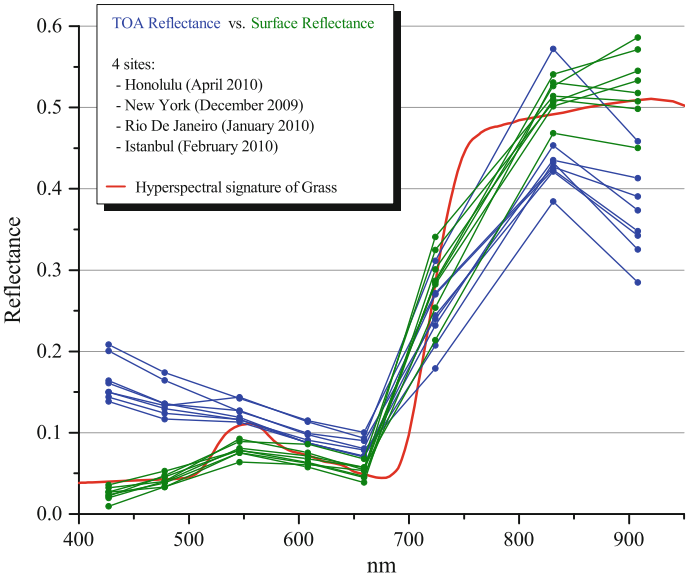


Fig. 3.11 Spectral signatures of grass in TOA (*blue lines*) and surface (*green lines*) reflectance values for four urbanized areas acquired by WorldView-2 (each *circle* represents the center wavelength of a spectral band). An additional hyperspectral signature of grass is reported with a red line for comparison. TOA reflectance signatures are consistently larger than the surface reflectance values at shorter wavelengths due to Rayleigh scattering. On the other hand, TOA reflectance values are smaller than the surface reflectance signatures at 949 nm due to water vapor absorption features

distortions with respect to the reference hyperspectral signatures, making it a suitable transformation for the analysis of large temporal data sets.

A comprehensive review of remote sensing quantities can be found in [92].

3.3.3 Accounting for Angular Variability

In the previous section, it was mentioned that the Sun-Earth-satellite geometry influences the measured at-sensor radiance, and therefore both DN and surface reflectance values. While the Earth-satellite geometry can be kept constant with nadir-looking sensors such as the Landsat series, commercial satellites operate with significantly varying geometries due to their agility in rapid re-targeting of the camera. This is especially important in the case of emergencies when the viewing geometries are opened up to maximize the collection opportunities.

In this section, geometrical effects are discussed in terms of surface anisotropy whose main mechanisms are illustrated in Sect. 3.3.3.1. In Sect. 3.3.3.2, two angular decomposition models are reviewed in detail.

3.3.3.1 Surface Anisotropy

A surface that reflects the incident energy equally in all directions is said to be Lambertian, and its reflectance is invariant with respect to illumination and viewing conditions. On the contrary, a surface is said to be anisotropic when its reflectance varies with respect to illumination or viewing geometries. These changes are driven by the optical and structural properties of the surface material [93]. In general, both natural and man-made surfaces show some degree of spectral anisotropy, and this behavior can be described by the bidirectional reflectance distribution function (BRDF) [94, 95].

The fundamental components of surface anisotropy as described in [96] are illustrated in Fig. 3.12:

- *surface scattering*, which can be observed when forward scattering elements are present and includes specular reflection
- radiative transfer-type *volumetric scattering*, which is due to the presence of finite scatterers, such as leaves of plants
- *geometrical-optical scattering*, which is produced by casting shadow and mutual obscuration of vertical surfaces.

Consequently, multi-angular acquisitions contain information about the physical structure and characteristics of the observed target. Depending on the study, surface anisotropy can be seen as a source of noise (e.g., when analyzing spectral signatures of time-series) or, alternatively, as a source of information in addition to the tempo-spectral dimensions [94].

To qualitatively show the effects of surface anisotropy across multi-temporal very high spatial resolution data sets, two examples are discussed as follows:

(1) The two images in Fig. 3.13 were acquired over Denver only five days apart in March 2008, but from opposite satellite azimuths, i.e., 150 and 339 degrees (see Table 3.3 for details). As shown, some surfaces appear brighter when observed from the backward direction, while others appear brighter when observed from

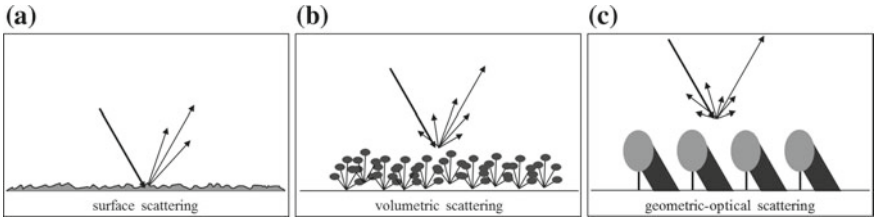


Fig. 3.12 Fundamental components of surface anisotropy: **a** surface, **b** radiative transfer-type volumetric, and **c** geometrical-optical scattering. Surface scattering includes specular reflection and can be observed when forward scattering elements are present; radiative transfer-type volumetric scattering is due to the presence of finite scatterers, such as leaves of plants; geometrical-optical scattering is produced by casting shadow and mutual obscuration of vertical surfaces

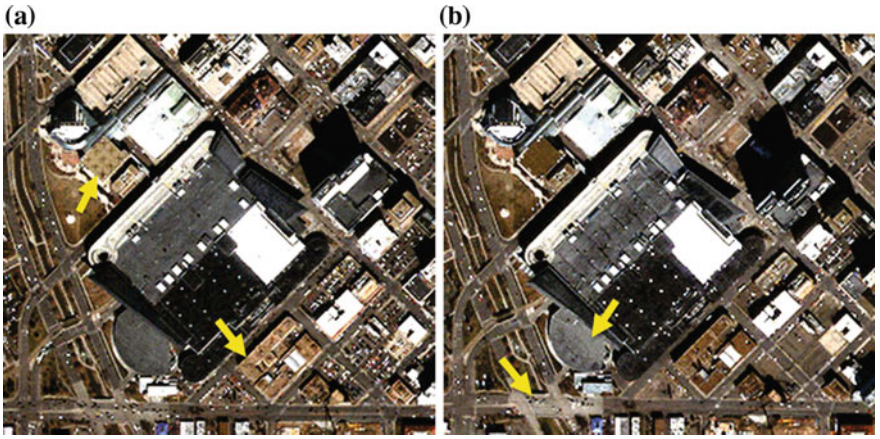


Fig. 3.13 Effects of surface anisotropy on two images acquired in March 2008 from opposite satellite azimuths (i.e., 150 and 339 degrees). As shown by the *yellow arrows*, (**a**) some surfaces appear brighter from the backward direction, while (**b**) others appear brighter from forward direction. The Sun illumination azimuth angle is approximately 158 degrees in both images (color figure online)

the forward direction. Therefore, image matching techniques that do not account for localized distributions will necessarily fail, possibly compromising the spectral meaning of the data.

(2) Figure 3.14 illustrates two domain representations of the Denver baseball stadium grass during only the summer acquisitions of the time-series. Fall or winter dates were not considered to avoid spectral variations related to the state of grass rather than to the different viewing conditions. This choice also minimizes the Sun declination, assuring a quasi-consistent source of illumination, in terms of position and intensity. Finally, it is worth mentioning that the grass of the stadium is kept uniform through the season according to baseball regulations. This means that this data set is suitable to illustrate the effects of different viewing conditions when illumination and target are unchanged. Specifically, the multi-temporal plot in Fig. 3.14a illustrates variations across the dates, with the near-infrared band ranging from about 0.44 to 0.62 just in the summer of 2005. The polar plot in Fig. 3.14b shows that these variations are consistent with the angular properties of the surface and the different viewing geometries (with elements of the previously discussed scattering mechanisms). In fact, observations in the backward direction consistently appear brighter than that in the forward direction.

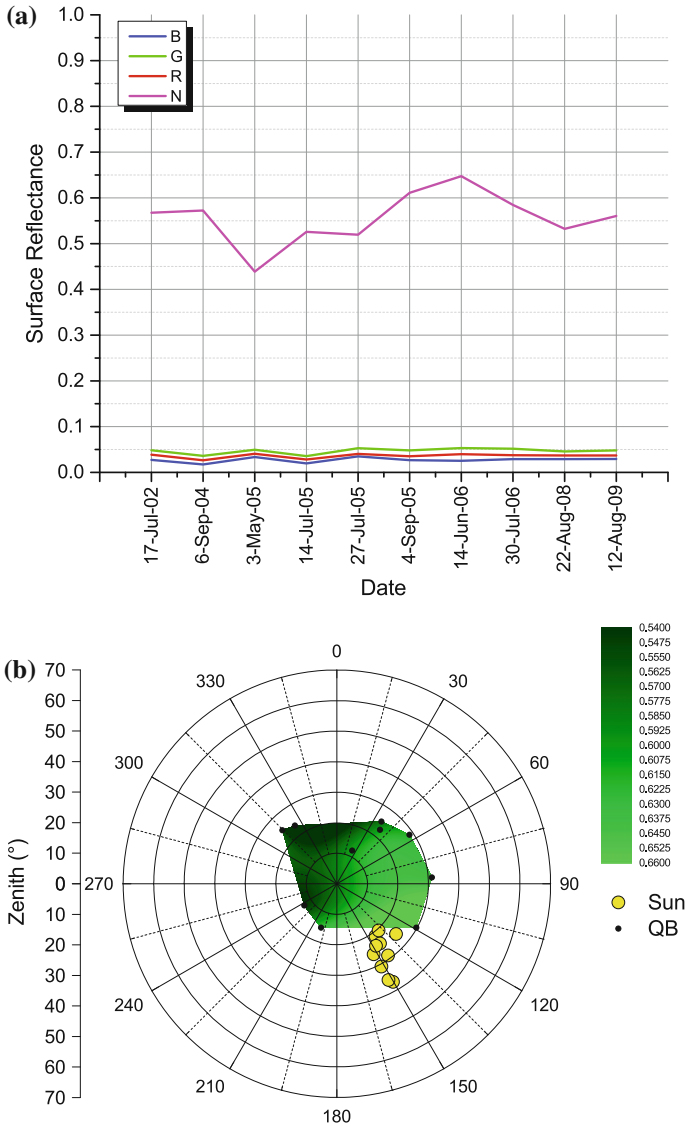


Fig. 3.14 Two domain representations of the Denver baseball stadium grass during only the summer acquisitions of the time-series: **a** multi-temporal signature and **b** near-infrared angular reflectance (with non-constant illumination geometry). The multi-temporal plot in **a** illustrates variations in the near-infrared band across the dates, ranging from about 0.44 to 0.65. The polar plot in **b** shows that these variations are consistent with the angular properties of the surface and the different viewing geometries. In fact, observations in the backward direction consistently appear brighter than that in the forward direction

3.3.3.2 Angular Decomposition Models

The increasing availability of multi-angular measurements from space-borne sensors, such as the Along-Track Scanning Radiometer-2 (ATSR-2), Polarization and Directionality of the Earth's Reflectances (POLDER), Multi-angle Imaging SpectroRadiometer (MISR), Compact High Resolution Imaging Spectrometer (CHRIS), WorldView-1, and WorldView-2, or aerial cameras, has brought new and unique opportunities to understand and exploit surface anisotropy. Several studies have proven over the years that medium and high spatial resolution angular data significantly improve the classification accuracy of various land covers as shown by Kimes et al. [97], Sandmeier and Deering [98], Chopping et al. [99], Armston et al. [100]. More recently, Su et al. [95], Verrelst et al. [101], Laurent et al. [102], and Koukal and Atzberger [103] showed that angular data are useful for vegetation mapping as it provides information that is not available in the spectral domain. Some studies addressing multi-angular observations were recently reviewed in a dedicated special issue [66].

A number of techniques have been proposed to characterize and efficiently use angular information. Two assumptions generally hold true [104]:

- Atmospheric effects should be removed
- A sufficient angular sampling should be available to provide robust retrievals

In particular, Koukal and Atzberger [103] found that angular observations should cover both the backward and forward scattering directions. Otherwise, models will over-fit the data, and the parameters retrieved will not represent the *true* anisotropy of the target.

Angular models can be classified as *physical* or *empirical* [105]. Physical models rely on first-principle physics and require a complete and comprehensive model parametrization (e.g., inputs such as surface roughness or complex refractive index), which is often very difficult to obtain. On the contrary, empirical models rely exclusively on measured angular values. A trade-off between these two techniques is represented by *semiempirical* models which incorporate measured data to elements of physics-based principles. Semiempirical models can be applied without any knowledge of the target complexity and composition, as they do not impose severe hypotheses about the nature and structure of the surface being modeled [93]. Among the various semiempirical models proposed, the kernel-driven Ross-Li [106] and the Rahman-Pinty-Verstraete (known as RPV) [93] are some of the most widely used. The former assumes that surface anisotropy can be described by the linear contribution of a set of kernels that describe the basic mechanisms of surface anisotropy, whereas the latter provides a representation of surface anisotropy by means of angular functions [94]. In particular, the Ross-Li model is the basis for the MODIS BRDF/albedo product, while the RPV model is used to generate the MISR BRDF/albedo product. It was found in [107, 108] that both models provide comparable results, with errors within 10% from observations [109].

Given the solar zenith angle, θ_s, the view zenith angle θ_v, and the relative view-solar azimuth angle ϕ, the Ross-Li model decomposes the observed angular surface

reflectance into three basic scatter mechanisms: isotropic, radiative transfer-type volumetric, and geometrical-optical. This model combines these elements as:

$$\rho_{Ross-Li}(\theta_s, \theta_v, \phi) = f_{iso} + f_{vol} \cdot K_{vol}(\theta_s, \theta_v, \phi) + \\ + f_{geo} \cdot K_{geo}(\theta_s, \theta_v, \phi) \tag{3.35}$$

where K_{vol} and K_{geo} are the volumetric and geometric scattering kernels, and f_{iso}, f_{vol}, and f_{geo} are the isotropic, volumetric, and geometrical kernel scaling factors. The isotropic scattering has no dependency on incidence or viewing angle, and therefore does not have a geometrically dependent kernel. The angular behavior of the volumetric kernel presents a minimum near the backward direction and bright limbs, while the angular behavior of geometrical kernel shows a maximum in the backward direction, where there are no shadows [104].

The RPV model decomposes the observed angular surface reflectance into three independent components, representing the amplitude ρ_0, the shape anisotropy k, and the asymmetry factor Θ, according to:

$$\rho_{RPV}(\theta_s, \theta_v, \phi) = \rho_0 \cdot \frac{\cos^{k-1}\theta_s \cos^{k-1}\theta_v}{(\cos\theta_s + \cos\theta_v)^{1-k}} \cdot \\ \cdot F(g, \Theta) \cdot H(G, \rho_0) \tag{3.36}$$

with:

$$F(g, \Theta) = \frac{1 - \Theta^2}{(1 + \Theta^2 + 2\Theta\cos g)^{3/2}} \tag{3.37}$$

$$H(G, \rho_0) = 1 + \frac{1 - \rho_0}{1 + G} \tag{3.38}$$

$$\cos g = \cos\theta_s \cos\theta_v + \sin\theta_s \sin\theta_v \cos\phi \tag{3.39}$$

$$G(\theta_s, \theta_v, \phi) = (\tan^2\theta_s + \tan^2\theta_v - 2\tan\theta_s \tan\theta_v \cos\phi)^{1/2} \tag{3.40}$$

The parameter $\rho_0 \in [0, 1]$ characterizes the intensity of the target, but it should not be confused with the single-scattering albedo or the *true* reflectance of the target, as it is independent of the angular variations. The parameter $k \in [0, 2]$ indicates the anisotropy of the target. Values of k smaller than 1.0 represent a bowl-shaped anisotropy pattern, where k increases with the view zenith angle. In contrast, values of k greater than 1.0 represent a bell-shaped anisotropy pattern, where k reaches its maximum at the nadiral view. A Lambertian surface is represented by the ideal case of $k = 0$. It is important to emphasize that k is influenced by the direction of the illumination with respect to the target. Therefore, the values of k in multi-temporal

data sets should be carefully interpreted not as an intrinsic property of the surface, i.e., k may vary as a function of the season [94]. Finally, the asymmetry factor $\Theta \in [-1, 1]$ controls the relative amount of forward, $\Theta \in (0, 1]$, and backward scattering, $\Theta \in [-1, 0)$.

A more applicative discussion about the retrieval of the RPV model (the triplet ρ_0, k, and Θ) is discussed in Sect. 3.3.4.4. The interested reader can also refer to [103].

3.3.4 Experimental Results

In the previous sections, it was discussed how surface reflectance provides a consistent feature space, and that an additional dimensionality can be exploited by including angular information. The results of four different experiments carried out over the Denver time-series are described in this section to support these concepts. The first two exercises focus only on the temporal aspects of the data set and on the advantages of working in the surface reflectance domain. In particular, Sect. 3.3.4.1 illustrates the analysis of the tempo-spectral variations of a rooftop, while Sect. 3.3.4.2 addresses the differences of two automated urban change detection methods over two sets of raw DNs and surface reflectance image-pairs. The other two exercises discuss the advantages of coupling temporal and angular information: In Sect. 3.3.4.3, the tempo-angular spectral signatures of an "unknown object" are analyzed with the aim of providing additional information about its characteristics, whereas Sect. 3.3.4.4 provides the results of a 22-class urban land cover exercise.

3.3.4.1 Analysis of Multi-temporal Spectral Signatures

The multi-temporal spectral signatures of a flat roof are shown in Fig. 3.15 in both DNs (normalized to 1.0 for sake of comparison) and surface reflectance values. These spectral signatures represent the average of all pixels over the roof. The aerosol optical depth (AOD), which is inversely related to the visibility (i.e., the lower the AOD, the higher the visibility), is also reported for completeness. The AOD is one of the outputs provided by the method described in [83, 84].

The DN curves in Fig. 3.15a show that the green values are consistently higher than the other spectral components, with very large variability in all bands through the years. This leads to the conclusion that the color of the roof is consistent with some shade of green, and no additional information can be reliably derived to explain, for example, the nature of the temporal behavior.

On the other hand, three well-defined, fairly stable, temporal regions can be identified from the surface reflectance values shown in Fig. 3.15b. Specifically, there is a relatively flat spectral plateau at about 0.35 reflectance, followed by a much darker response between the winter of 2007 and the spring of 2008, and then by a highly reflecting region in subsequent periods to the end of the time-series (about 0.70 reflectance). These relatively flat temporal regions, and their sharp transitions, may

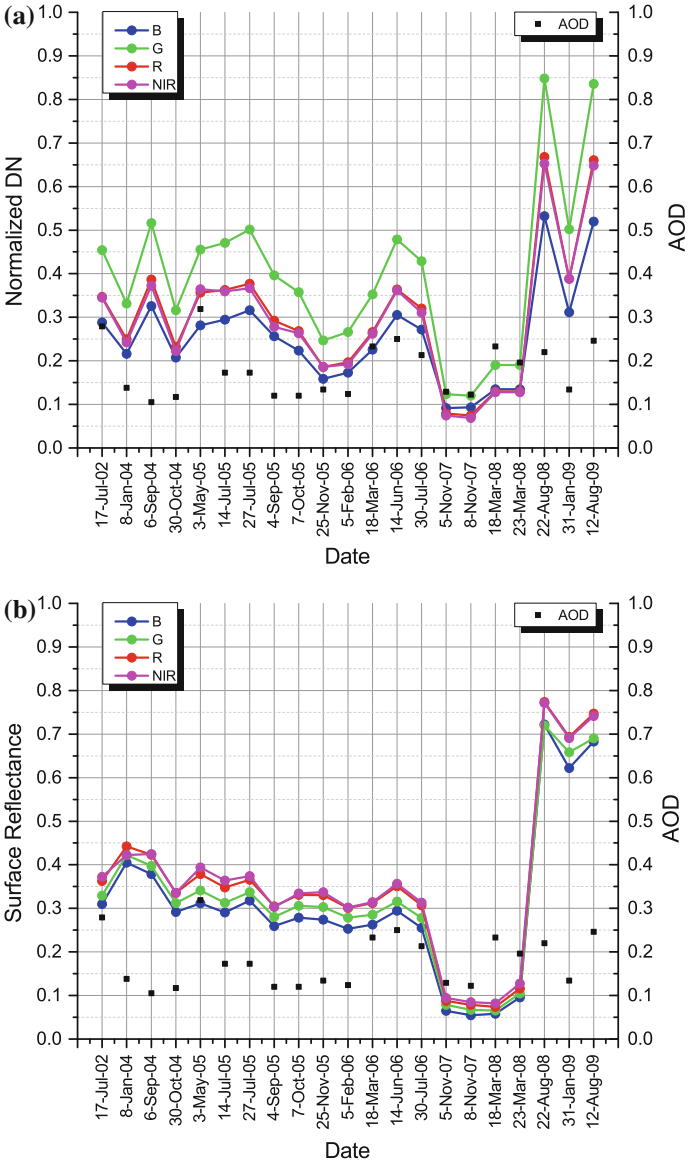


Fig. 3.15 Multi-temporal spectral signatures in **a** normalized DNs and **b** surface reflectance of a remodeled flat roof. The temporal transition is illustrated in the image sequence in **c** which shows that the roof was being remodeled between the end of 2007 and the beginning of 2008. The date-to-date variability of the DN temporal signatures is related to the AOD values with a Pearson correlation coefficient of 0.427, which is significant at the 0.05 confidence level. However, the date-to-date variability of the surface reflectance temporal signatures is correlated with the AOD values with a Pearson coefficient of 0.097, which is no longer significant at the 0.05 level

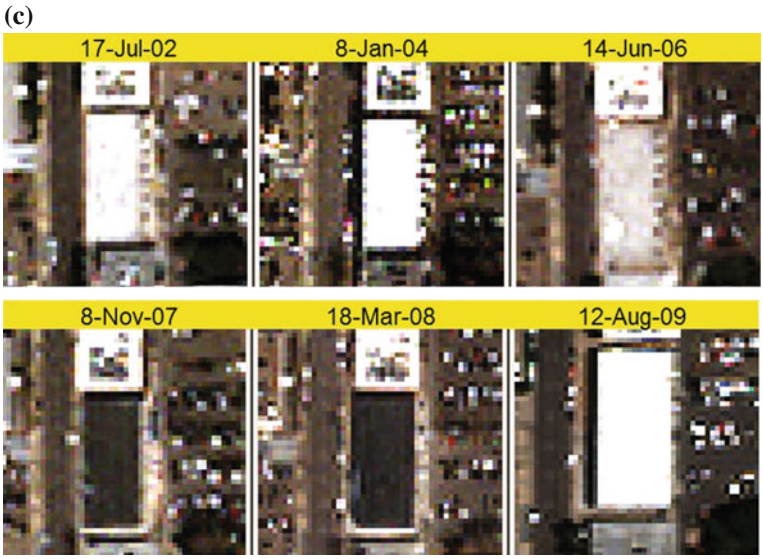


Fig. 3.15 (continued)

indicate that the roof being investigated went through renovation during the years. Also, there is a small difference between the bands at any date in the sequence (about 0.05 or less in absolute terms), suggesting that the roof is consistent with three different shades of gray, from a medium tone up to 2007, to a very bright shade after 2008, with a dark response in between.

The temporal behavior can be understood looking at Fig. 3.15c, which shows that the roof was being remodeled between the end of 2007 and the beginning of 2008. Therefore, a simple analysis of surface reflectance signatures indicates that a physical input space can be extremely helpful in understanding the temporal variability and its sharp transitions, whereas a DN-based analysis failed to provide consistent information. It is also worth mentioning that the date-to-date variability of the DN temporal signatures in Fig. 3.15a is related to the AOD values with a Pearson correlation coefficient of 0.427, which is significant at the 0.05 confidence level. However, the date-to-date variability of the surface reflectance temporal signatures in Fig. 3.15b is correlated with the AOD values with a Pearson coefficient of 0.097, which is no longer significant at the 0.05 level. This emphasizes the fact that atmospheric conditions significantly affect the DN input domain, whereas the date-to-date variability of the surface reflectance curves can be attributed to the anisotropic nature of the target. This is further discussed in Sect. 3.3.4.3, which illustrates the analysis of multi-temporal spectral signatures coupled to their angular components.

3.3.4.2 Automated Urban Change Detection

Several methodologies have been developed during the years in the context of automated change detection. Two widely used approaches are based on change vector analysis (CVA) and principal component analysis (PCA). In the former, pixels are represented by their vectors in the feature space and the changes are derived as the difference of the feature vectors between the images [110–113]. In the latter, the first principal component (which corresponds to the largest eigenvalue) reflects the unchanged parts of a set of images, whereas changes can be depicted from the components corresponding to smaller eigenvalues [114–117]. See Chap. 8 for more details on change detection.

Two sets of raw DNs and surface reflectance image-pairs are used to qualitatively and quantitatively investigate the differences of non-physical and physical domains for change detection studies. The image-pairs correspond to a subset of the entire scene, where several changes occurred between July 2002 and August 2008 (shown in Fig. 3.16a, b, respectively). These changes include the construction and demolition of several large buildings in the industrial area, and the remodeling of various roofs in the residential district. The two dates were selected as the corresponding images were acquired with similar viewing geometries and during similar time of year, minimizing both stereoscopics and seasonal effects (which are not the focus of this analysis). The near-infrared band was not considered to filter out changes in vegetation cover, which are considered not relevant in this analysis.

Figure 3.16 shows the CVA and PCA change detection results derived using DN counts as input (Fig. 3.16c and e, respectively) and using surface reflectance values (Fig. 3.16d and f, respectively), where the magnitude of change is normalized to the interval $[-1.0, +1.0]$. Extensive false alarms are visible, especially over paved surfaces and rooftops, in both change maps that were produced from DN counts. On the other hand, the change maps produced from surface reflectance values clearly identify the changes due to the construction of new buildings (shown in cold colors) and the demolition of older structures (shown in warm colors).

Despite the simplicity of both methodologies, the improvement obtained from the use of surface reflectance data is evident. This is quantitatively illustrated in Table 3.4, which reports the overall accuracy (OA), false alarms (FA), and missed alarms (MA) for the four cases and six thresholds levels. As expected, the larger the threshold, the smaller the FA and the larger the MA rates. The goal of automated, semiautomated, and manual thresholding techniques is to find a value that maximizes the OA and minimizes both the FA and the MA. For example, small thresholds can be used when there is no bias between the average values of the two acquisitions, whereas larger thresholds are necessary to account for differences in the data distributions. Table 3.4 shows that low thresholds (i.e., 0.05 and 0.10) provide high accuracy with low FA and MA values only for the surface reflectance cases. It is interesting to point out that for both CVA DN and PCA DN, only a large threshold (0.25 or 0.30) can result in an acceptable FA rate, but at the price of a MA rate well above 30%.

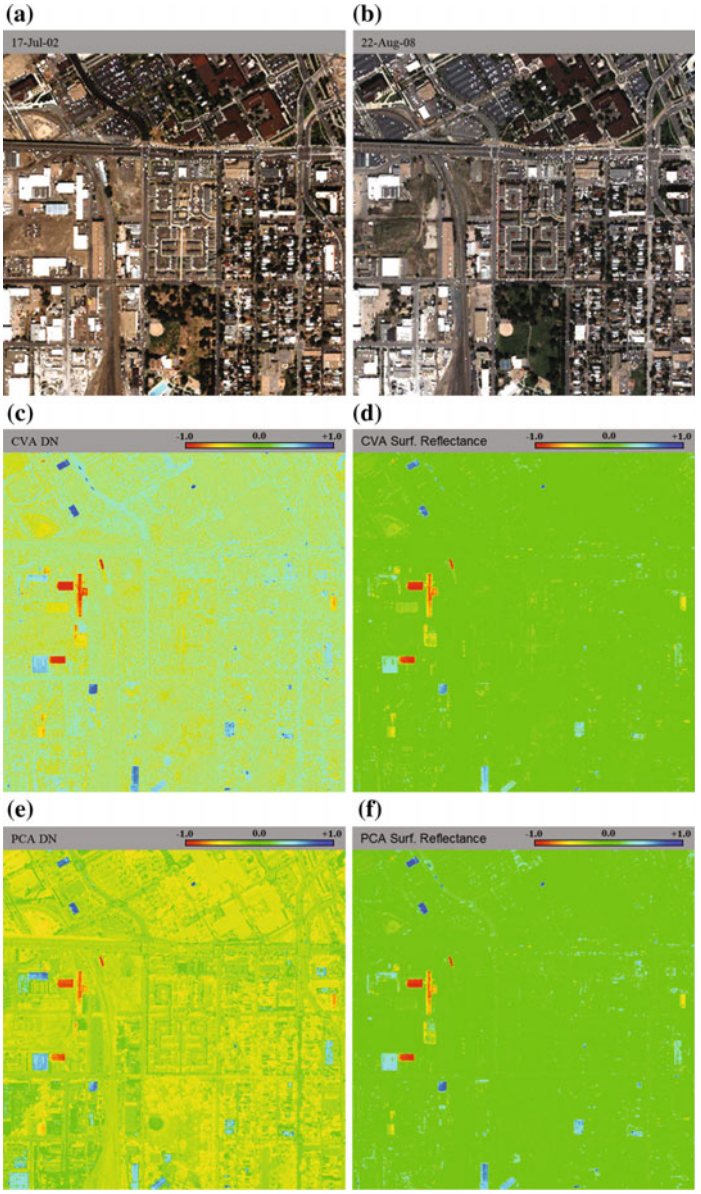


Fig. 3.16 The two scenes used to generate the change detection maps were acquired on **a** July 2002 and **b** August 2008. The change detection maps in **c** and **e** were derived from DN counts as input using the CVA and PCA approaches, respectively, whereas the change maps in **d** and **f** were derived from surface reflectance values using the CVA and PCA approaches, respectively. Extensive false alarms are visible, especially over paved surfaces and rooftops, in both change maps that were produced from DN counts. On the other hand, the change maps produced from surface reflectance values clearly identify the changes due to the construction of new buildings (shown in cold colors) and the demolition of older structures (shown in warm colors)

Table 3.4 CVA and PCA change detection results

Threshold	OA (%)			
	CVA DN	PCA DN	CVA SURF	PCA SURF
0.05	2.67	2.99	78.23	78.12
0.10	4.68	4.99	94.59	94.21
0.15	12.14	9.09	97.91	97.67
0.20	47.29	20.58	98.86	98.72
0.25	84.03	48.72	99.09	99.10
0.30	94.59	80.04	99.05	99.12
Threshold	FA (%)			
	CVA DN	PCA DN	CVA SURF	PCA SURF
0.05	98.85	98.56	22.00	22.15
0.10	96.72	96.49	5.26	5.70
0.15	89.08	92.25	1.78	2.06
0.20	53.26	80.49	0.68	0.88
0.25	15.86	51.74	0.24	0.35
0.30	4.98	19.57	0.12	0.16
Threshold	MA (%)			
	CVA DN	PCA DN	CVA SURF	PCA SURF
0.05	3.99	2.06	7.69	0.99
0.10	9.70	4.38	14.94	11.29
0.15	13.61	8.89	20.78	18.72
0.20	18.94	14.21	29.15	25.12
0.25	22.58	23.01	41.52	33.83
0.30	30.87	43.50	51.70	44.61

3.3.4.3 Combined Analysis of Multi-temporal and Multi-angular Spectral Signatures

Similar to Fig. 3.15a, the AOD and the multi-temporal spectral signature in normalized DNs of an "unknown object" are illustrated in Fig. 3.17a. With this plot, it is very difficult to extract useful information on the nature of the surface being investigated. For example, there are evident variations in the near-infrared band between winter and summer acquisitions (from about 0.16 to 0.36), which may be indicative of a natural surface (such as vegetation). However, it is not possible to guess with confidence the color of the object as the DN values are influenced by several factors as discussed previously. In this particular case, one can believe that the object may be yellow, as the red and green bands are very close in amplitude to each other. Also

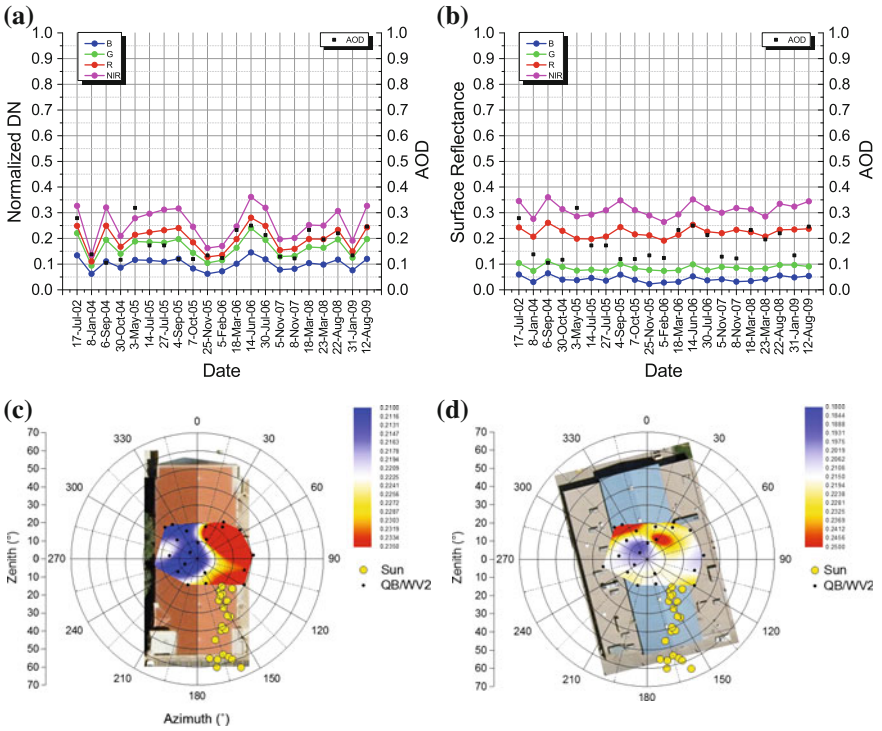


Fig. 3.17 Multi-temporal spectral signatures in **a** normalized DNs and **b** surface reflectance for the pitched brick-red roof shown in background in **c** which illustrates its angular surface reflectance of the *red band* (with natural, non-constant, illumination). The *yellow circles* represent the Sun locations throughout the time-series in zenith and azimuth angles, whereas the black dots correspond to the satellite positions. An additional angular plot for a *blue* pitched roof (shown in background) is illustrated in **d** (color figure online)

in this case, it is worth mentioning that the date-to-date variability is related to the AOD values with a Pearson coefficient of 0.538, which is significant at the 0.05 level.

On the other hand, from Fig. 3.17b, which represents the same plot as in Fig. 3.17a, but in the surface reflectance domain, it is possible to deduce that the "unknown object" has a smaller seasonal variability in the near-infrared band (from 0.26 to 0.35) with peaks not necessarily corresponding to the warmer months (see, for example, the image acquired on January 2009), which can lead to the conclusion that the object may not be a natural surface. Further, the "unknown object" is red, as the green and the blue spectral signatures are constantly below the red one. Therefore, using only a multi-temporal analysis, it is possible to assess (with some degree of uncertainty) that the "unknown object" is consistent with the assumptions of a red object, possibly man-made. In this case, the date-to-date variability is related to the AOD values with a Pearson coefficient of −0.013, which is not significant at the 0.05 level.

Figure 3.17c illustrates the angular surface reflectance of the red band for the "unknown object" (with natural, non-constant, illumination). Similar to Fig. 3.14, the yellow circles represent the Sun locations throughout the time-series in zenith and azimuth angles, whereas the black dots correspond to the satellite positions. This plot is particularly useful to extract additional information about the structure of the surface being investigated. In fact, it clearly shows that the "unknown object" has a consistent bimodal reflectance distribution (brighter on the acquisitions from East, and darker from West), which is consistent with the structure of a pitched object (where only the side on the East is directly illuminated by the Sun). For the sake of completeness, Fig. 3.17c also shows, in background, the image of the "unknown object," which is a brick-red pitched roof.

Another example of angular surface reflectance for a non-constant illumination geometry is reported in Fig. 3.17d, which represents a blue roof, as shown in the background image. In particular, the angular surface reflectance shows a fairly consistent plateau of about 0.18 in the azimuthal southern part of the plot, and two areas with higher reflectance (about 0.25) in the Sun specular reflection. From this, it is possible to understand that the roof is also pitched (because of the two disjoint areas of higher reflectance).

It is worth mentioning that, as for Sect. 3.3.4.1, the analysis of the spectral signatures represents the average of all pixels over each of the two roofs.

3.3.4.4 Urban Land Cover Classification

In order to quantitatively investigate the benefits of surface reflectance and angular decompositions to improve urban land cover classification of image time-series, 22 *non-common* classes of interest were selected from the Denver data set. These classes include different kinds of grass, water, soil, paved surfaces, and pitched or flat roofs as shown in Table 3.5. The angular decomposition used for this experiment is RPV [94].

Three representation domains of the time-series data set, DN, surface reflectance, and RPV, were used in three independent classification experiments. The 21 images were randomly sampled by 101 different cross-validation runs, where 11 dates were retained for training and the remaining for validation. These independent sets were used to create two, distinct RPV models (one for training and one for validation). A random forest model with 100 trees was generated for each cross-validation run and each input domain [90]. It is worth noting that the models for the DN and surface reflectance domains assumed an input space composed by four features, derived from the four QuickBird bands, whereas the RPV input space was composed of 3×4 features, as the RPV decomposition provides ρ_0, k, and Θ for each spectral band.

The RPV decomposition was implemented by fitting the model to different lookup tables (LUTs), one for each acquisition, as the illumination and viewing geometries change for each date. In particular, the RPV parameters were sampled for the generation of the LUTs according to the following scheme:

Table 3.5 Classes of interest

Natural surfaces	Man-made surfaces
Grass type 1 (stadium)	Parking lot type 1 (asphalt, bright)
Grass type 2 (golf-fairway)	Parking lot type 2 (asphalt, dark)
Grass type 3 (golf-green)	Roof type 1 (brown, dark, flat)
Grass type 4 (park 1)	Roof type 2 (brown, bright, flat)
Grass type 5 (park 2)	Roof type 3 (concrete, gray, flat)
Tree	Roof type 4 (concrete, bright, flat)
Water type 1 (lake)	Roof type 5 (concrete, dark, flat)
Water type 2 (river)	Roof type 6 (red, brick, pitched)
Soil type 1	Roof type 7 (blue, pitched)
Soil type 2 (playground)	Roof type 8 (metal, pitched)
Soil type 3 (baseball field)	Shadow (of various man-made materials)

- $\rho_0 \in [0.0 : 0.01 : 1.0]$
- $k \in [0.0 : 0.01 : 2.0]$
- $\Theta \in [-1.0 : 0.01 : 1.0]$

Successively, for each class of interest, the 11 LUT entries that provided the smallest root-mean-square error between the retrieved values and the one measured by the sensor were retained, and their average value was considered as the solution of ρ_0, k, and Θ.

Figure 3.18 illustrates the retrieved near-infrared values of ρ_0, k, and Θ for the five types of grass. For the sake of completeness, the near-infrared values in DNs and surface reflectance domains are also reported. As shown, it is not possible to accurately define more than one single cluster of grass with DN values due to the large degree of overlap between the five distributions. On the other hand, the surface reflectance domain reduces the intra-class variability of the different types, allowing the identification of two separate clusters of healthy and less healthy grass (*grass type 1–2–3* and *grass type 4–5*, respectively) with a threshold of 0.48. Finally, the triplet ρ_0, k, and Θ allows the discrimination of all five grass types with fairly good confidence (as discussed later in more detail). In particular, the intensity ρ_0 contains similar information as that provided by the surface reflectance domain, making it possible to cluster healthy and less healthy grass with a threshold of 0.31. The anisotropy factor k has values smaller than 1.0 for all grass types but *grass type 1*, representing a bowl-shaped pattern as found in [118]. The class *grass type 1* shows an anisotropy factor consistent with a bell-shaped pattern which can be explained by the renovations of the stadium field during winter acquisitions, when the grassy surface is generally converted to bare soil [119]. This also explains the larger variability of this class with respect to the others in all domains. Further, as discussed earlier, the anisotropy factor k is more sensitive to temporal changes [94] as shown by the larger intra-class variability compared to ρ_0 and Θ. Overall, k differentiates *grass type 1* from *grass type 4*, and the other grass types (with thresholds 1.0 and 0.8, respectively).

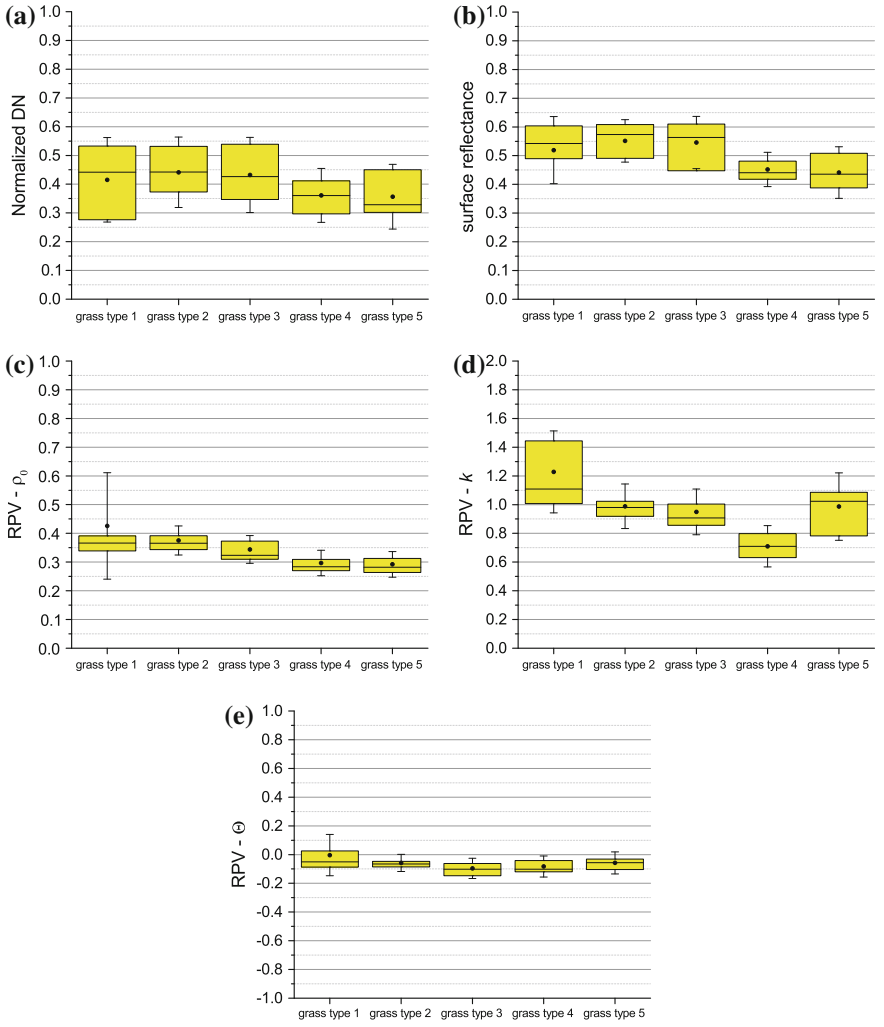


Fig. 3.18 The near-infrared values in **a** DNs and **b** surface reflectance domains for the five types of grass are reported in Table 3.5. The RPV retrieved near-infrared values of ρ_0, k, and Θ are shown in **c**, **d**, and **e**, respectively. As shown, it is not possible to accurately define more than one single cluster of grass with DN values due to the large degree of overlap between the five distributions. The surface reflectance domain reduces the intra-variability of the different types, allowing the identification of two separate clusters of healthy and less healthy grass (*grass type 1–2–3* and *grass type 4–5*, respectively). The triplet ρ_0, k, and Θ allows the discrimination of all five grass types with fairly good confidence as reported by the F_1 measure in Table 3.6

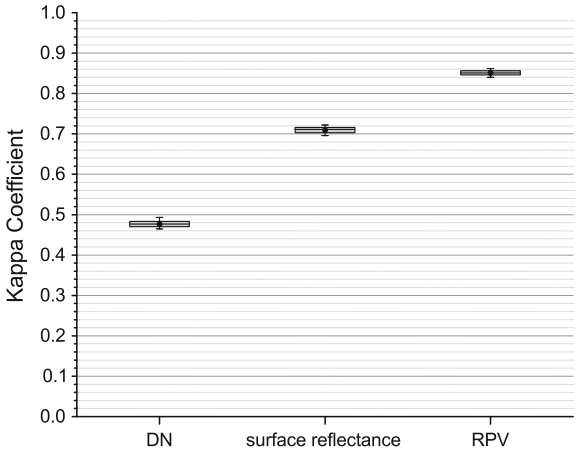


Fig. 3.19 Box plots of the 101 cross-validation runs for the three independent experiments using DN, surface reflectance, and RPV values. The three results are statistically significant according to the McNemar's test

The asymmetry factor Θ is systematically below 0.0 for all types of grass, which is consistent with the backward scattering behavior as also evinced from Fig. 3.14. The use of Θ allows the discrimination of *grass type 2* from *grass type 3* (with a threshold of -0.09).

Figure 3.19 illustrates the classification accuracies (in terms of Kappa coefficient) for the three domains. The box plots account for the variability of both the different random forest model initializations and training/testing sets of the 101 cross-validation runs. The three results are statistically significant according to the McNemar's test. The average accuracy when DN values are used as input is about 0.477. This result is quite in line with initial expectations as some of the classes are spectrally similar to each other, such as the five type of grass, or they may require knowledge of the structure being analyzed, such as the pitched roof classes. The F_1 measure of the various targets is reported in Table 3.6. As shown, the grass classes are not differentiated, with a maximum F_1 just slightly above 0.3. In general, the remaining classes have an F_1 value around 0.5. However, it is interesting to note how *roof type 6* and *roof type 7* are actually quite well discriminated with an F_1 value of over 0.85. This may be due to the very distinct spectral components of these two roofs (which appeared red and blue in the visible bands, respectively) with respect to the other classes. An improvement over the DN case of about 49.1% is achieved by considering surface reflectance values, corresponding to a Kappa coefficient of 0.711. Roughly, all classes show better accuracies even though the grass and soil clusters are still not well differentiated, which can be explained by the spectral similarity of these targets. Finally, a Kappa coefficient of 0.851 is achieved when surface reflectance is combined to the RPV decomposition of the time-series, corresponding to an improvement of 78.4% over the base case of DNs. The F_1 values of the grass and soil classes are all well above 0.70, which can be seen as a satisfactory result due to the complex nature of this task.

Table 3.6 F_1 measure of the 22 classes of interest

Class	DN	Surface reflectance	RPV
Grass type 1 (stadium)	0.324	0.682	0.923
Grass type 2 (golf-fairway)	0.182	0.524	0.808
Grass type 3 (golf-green)	0.350	0.585	0.740
Grass type 4 (park 1)	0.213	0.341	0.769
Grass type 5 (park 2)	0.095	0.341	0.714
Tree	0.650	0.829	0.976
Water type 1 (lake)	0.372	0.722	0.932
Water type 2 (river)	0.500	0.683	0.935
Soil type 1	0.488	0.432	0.701
Soil type 2 (playground)	0.489	0.681	0.920
Soil type 3 (baseball field)	0.537	0.632	0.884
Parking lot type 1 (asphalt, bright)	0.378	0.667	0.615
Parking lot type 2 (asphalt, dark)	0.458	0.844	0.905
Roof type 1 (brown, dark, flat)	0.591	0.889	0.926
Roof type 2 (brown, bright, flat)	0.421	0.694	0.889
Roof type 3 (concrete, gray, flat)	0.444	0.976	0.981
Roof type 4 (concrete, bright, flat)	0.533	0.682	0.737
Roof type 5 (concrete, dark, flat)	0.512	0.773	0.651
Roof type 6 (red, brick, pitched)	0.857	0.976	0.976
Roof type 7 (blue, pitched)	0.895	0.901	0.904
Roof type 8 (metal, pitched)	0.585	0.718	0.884
Shadow (of various man-made materials)	0.558	0.870	0.917

3.4 Conclusions

The protocol for image information mining discussed in the first part of this chapter was designed to offer a high level of interactivity and to be adaptive to the image information content. The mining methodology consists of an interactive selection of positive and negative samples from the image space that are mapped into the feature space through the Max-Tree or Alpha-Tree structure. Both structures organize the image information content hierarchically and offer efficient means for component labeling and attribution/feature extraction. Feature spaces can be computed directly from the either of the two trees rapidly. The kd-Tree manages them by organizing the exported features in custom and application suitable ways from which accurate classifications can be computed. Typical performance figures for this protocol measured for the hardware reported in Sect. 3.2.6 and operated on a gigapixel VHsR satellite image are as follows: 3 min. for computing the Max-Tree, 3 min. for computing the kd-Tree, and interactive querying of less than 10 s.

The availability of sub-meter optical imagery regularly acquired over the same geographical region has proved to be effective in a large number of applications, from precision agriculture, to disaster management, to urban planning. Despite the progress in space technology in operating more sophisticated sensors and the large amount of data made available, very little research addresses the advantages and challenges of multi-temporal and multi-angular optical very high spatial resolution space-borne imagery. The second part of this chapter illustrated not only that physical quantities are necessary to consistently and efficiently analyze these kind of data sets, but also that the angular information of the acquisitions should not be neglected, as unique, additional features can be derived from it. More importantly, the temporal and angular components should always be simultaneously considered as some of the radiometric differences in the time-series (the so called *data set shift* in the machine learning terminology) may often be leveraged or accounted for by understanding the physics of the acquisitions. In this sense, it was shown that atmospheric and geometric properties of the acquisitions largely affect the image values, and significant correlation to AOD was found for the case of DN counts. Results of a 22-class urban land cover experiment showed that an improvement of 0.374 in terms of Kappa coefficient can be achieved over the base case of DNs when surface reflectance values are combined to the angular decomposition of the time-series.

References

1. Serpico, S., Bruzzone, L., Roli, F.: An experimental comparison of neural and statistical non-parametric algorithms for supervised classification of remote-sensing images. Pattern Recognit. Lett. **17**(13), 1331–1341 (1996)
2. Seidel, K., Schroder, M., Rehrauer, H., Datcu, M.: Meta features for remote sensing image content indexing. In: IEEE International Conference on Geoscience and Remote Sensing Symposium, vol. 2 (1998)
3. Inglada, J.: Automatic recognition of man-made objects in high resolution optical remote sensing images by svm classification of geometric image features. ISPRS J. Photogramm. Remote Sens. **62**(3), 3 (2007)
4. Datcu, M., Seidel, K.: Image information mining: exploration of image content in large archives. In: IEEE Proceedings of Aerospace Conference, vol. 3, pp. 253–264 (2000)
5. Demir, B., Persello, C., Bruzzone, L.: Batch-mode active-learning methods for the interactive classification of remote sensing images. IEEE Trans. Geosci. Remote Sens. **49**(3), 1014–1031 (2011)
6. Aksoy, S., Koperski, K., Tusk, C., Marchisio, G., Tilton, J.C.: Learning bayesian classifiers for scene classification with a visual grammar. IEEE Trans. Geosci. Remote Sens. **43**(3), 581–589 (2005)
7. Costache, M., Maitre, H., Datcu, M.: Categorization based relevance feedback search engine for earth observation images repositories. In: IEEE International Conference on Geoscience and Remote Sensing Symposium, pp. 13–16 (2006)
8. Schroder, M., Rehrauer, H., Seidel, K., Datcu, M.: Interactive learning and probabilistic retrieval in remote sensing image archives. IEEE Trans. Geosci. Remote Sens. **38**(5), 2288–2298 (2000)
9. Daschiel, H., Datcu, M.: Design and evaluation of human-machine communication for image information mining. IEEE Trans. Multimedia **7**(6), 1036–1046 (2005)

10. Blanchart, P., Ferecatu, M., Datcu, M.: Mining large satellite image repositories using semi-supervised methods. In: IEEE International Conference of Geoscience and Remote Sensing Symposium. Vancouver, Canada (2011)
11. Gueguen, L., Pesaresi, M., Soille, P.: An interactive image mining tool handling gigapixel images. In: IEEE International Conference of Geoscience and Remote Sensing Symposium, pp. 1581–1584 (2011)
12. Gueguen, L., Datcu, M.: Image time-series data mining based on the information-bottleneck principle. IEEE Trans. Geosci. Remote Sens. **45**(4), 827–838 (2007)
13. Kyrgyzov, I.O., Kyrgyzov, O.O., Maître, H., Campedel, M.: Kernel mdl to determine the number of clusters. In: Proceedings of 5th International Conference on Machine Learning and Data Mining in Pattern Recognition MLDM '07, pp. 203–217. Springer, Berlin (2007)
14. Baatz, M., Schape, A.: Multiresolution segmentation: an optimization approach for high quality multi-scale image segmentation. J. Photogramm. Remote Sens. **58**(3–4), 12–23 (2000)
15. Tilton, J., Marchisio, G., Koperski, K., Datcu, M.: Image information mining utilizing hierarchical segmentation. In: IEEE International Conference Geoscience and Remote Sensing Symposium, vol. 2, pp. 24–28 (2002)
16. Ouzounis, G.K., Soille, P.: The Alpha-Tree algorithm. JRC Technical reports, European Commission, Joint Research Centre, Institute for the Protection and Security of the Citizen (2012)
17. Ouzounis, G.K., Pesaresi, M., Soille, P.: Differential area profiles: decomposition properties and efficient computation. IEEE Trans. Pattern Anal. Mach. Intell. **34**(8), 1533–1548 (2012)
18. Pesaresi, M., Ouzounis, G.K., Gueguen, L.: A new compact representation of morphological profiles: report on first massive VHR image processing at the JRC. In: SPIE Proceedings, Baltimore, USA, 24–26 October 2012, vol. 8390, pp. 839025–839025–6 (2012)
19. Passat, N., Naegel, B., Rousseau, F., Koob, M., Dietermann, J.: Interactive segmentation based on component trees. Pattern Recognit. **44**, 2539–2554 (2011)
20. Najman, L.: On the equivalence between hierarchical segmentations and ultrametric watersheds. J. Math. Imaging Vis. **40**(3), 231–247 (2011)
21. Soille, P., Najman, L.: On morphological hierarchical representations for image processing and spatial data clustering. In: Köthe, U., Montanvert, A., Soille, P. (eds.) Applications of Discrete Geometry and Mathematical Morphology. Lecture Notes in Computer Science, vol. 7346, pp. 43–67. Springer, Berlin (2012)
22. Ehrlich, D., Kemper, T., Blaes, X., Soille, P.: Extracting building stock information from optical satellite imagery for mapping earthquake exposure and its vulnerability. Nat. Hazards **68**(1), 79–95 (2013)
23. Salembier, P., Oliveras, A., Garrido, L.: Anti-extensive connected operators for image and sequence processing. IEEE Trans. Image Process. **7**(4), 555–570 (1998)
24. Jones, R.: Connected filtering and segmentation using component trees. Comput. Vis. Image Underst. **75**(3), 215–228 (1999)
25. Westenberg, M.A., Roerdink, J.B.T.M., Wilkinson, M.H.F.: Volumetric attribute filtering and interactive visualization using the max-tree representation. IEEE Trans. Image Process. **16**(12), 2943–2952 (2007)
26. Urbach, E.R., Roerdink, J.B.T.M., Wilkinson, M.H.F.: Connected shape-size pattern spectra for rotation and scale-invariant classification of gray-scale images. IEEE Trans. Pattern Anal. Mach. Intell. **29**(2), 272–285 (2007)
27. Ouzounis, G.K., Soille, P.: Pattern spectra from partition pyramids and hierarchies. In: Soille, P., Pesaresi, M., Ouzounis, G.K. (eds.) (ISMM) 2011 Proceedings of 10th International Symposium Mathematical Morphology and its Applications to Image and Signal Processing. Lecture Notes in Computer Science, vol. 6671, pp. 108–119. Springer (2011)
28. Ouzounis, G.K., Syrris, V., Gueguen, L., Soille, P.: The switchboard platform for interactive image information mining. In: Proceedings ESA-EUSC-JRC 8th Conference Image Information Mining, Oberpfaffenhofen, Germany, 24–26 October 2012, pp. 26–30 (2002)
29. Ouzounis, G.K.: Design principles for interactive image information mining systems. In: Proceedings of 11th International Conference Pattern Recognition and Image Analysis, Samara, Russia, pp. 64–67 (2013)

30. Gueguen, L., Ouzounis, G.K.: Hierarchical data representation structures for interactive image information mining. Int. J. Image Data Fusion **3**(3), 221–241 (2012)
31. Ouzounis, G.K., Gueguen, L.: Interactive collection of training samples from the max-tree structure. In: Proceedings 18th IEEE International Conference of Image Processing (ICIP) 2011, Brussels, Belgium, pp. 1449–1452 (2011)
32. Serra, J. (ed.): Image Analysis and Mathematical Morphology. II: Theoretical Advances. Academic Press, London (1988)
33. Braga-Neto, U.M., Goutsias, J.: A theoretical tour of connectivity in image processing and analysis. J. Math. Imaging Vis. **19**, 5–31 (2003)
34. Braga-Neto, U.M., Goutsias, J.: Grayscale level connectivity: theory and applications. IEEE Trans. Image Process. **13**(12), 1567–1580 (2004)
35. Serra, J.: Connectivity on complete lattices. J. Math. Imaging Vis. **9**, 231–251 (1998)
36. Serra, J.: Connections for sets and functions. Fundamenta Informaticae **41**, 147–186 (2000)
37. Wilkinson, M.H.F., Ouzounis, G.K.: Advances in connectivity and connected attribute filters. In: Hawkes, P.W. (ed.) Advances in Imaging and Electron Physics, vol. 163, pp. 219–222. Elsevier (2010)
38. Wilkinson, M.H.F.: An axiomatic approach to hyperconnectivity. In: Wilkinson, M.H.F., Roerdink, J.B.T.M. (eds.) Proceedings of 9th International Symposium Mathematical Morphology and its Applications to Signal and Image Processing, (ISMM) 2009. Lecture Notes in Computer Science, vol. 5720, pp. 35–46. Springer, Berlin (2009)
39. Ouzounis, G.K., Wilkinson, M.H.F.: Hyperconnected attribute filters based on k-flat zones. IEEE Trans. Pattern Anal. Mach. Intell. **33**(2), 224–239 (2011)
40. Wilkinson, M.H.F.: Attribute-space connectivity and connected filters. Image Vis. Comput. **25**, 426–435 (2007)
41. Wilkinson, M.H.F.: Hyperconnectivity, attribute-space connectivity and path openings: theoretical relationships. In: Wilkinson, M.H.F., Roerdink, J.B.T.M. (eds.) Proceedings of 9th International Symposium Mathematical Morphology and its Applications to Signal and Image Processing, (ISMM) 2009. Lecture Notes in Computer Science, vol. 5720, pp. 47–58. Springer, Berlin (2009)
42. Soille, P.: Constrained connectivity for hierarchical image decomposition and simplification. IEEE Trans. Pattern Anal. Mach. Intell. **30**(7), 1132–1145 (2008)
43. Ronse, C.: Openings: Main properties, and how to construct them. Technical report, Philips Research Laboratory Brussels (1990)
44. Ronse, C.: Set-theoretical algebraic approaches to connectivity in continuous or digital spaces. J. Math. Imaging Vis. **8**, 41–58 (1998)
45. Breen, E.J., Jones, R.: Attribute openings, thinnings and granulometries. Comput. Vis. Image Underst. **64**(3), 377–389 (1996)
46. Heijmans, H.J.A.M.: Composing morphological filters. IEEE Trans. Image Process. **6**(5), 713–723 (1997)
47. Heijmans, H.J.A.M.: Connected morphological operators for binary images. Comput. Vis. Image Underst. **73**(1), 99–120 (1999)
48. Salembier, P., Serra, J.: Flat zones filtering, connected operators, and filters by reconstruction. IEEE Trans. Image Process. **4**(8), 1153–1160 (1995)
49. Hu, M.K.: Visual pattern recognition by moment invariants. IRE Trans. Inf. Theory **IT-8**, 179–187 (1962)
50. Gueguen, L., Ouzounis, G.K.: Tree based representations for fast information mining from VHR images. In: Proceedings of ESA-EUSC-JRC 8th Conference Image Information Mining, Oberpfaffenhofen, Germany, 24–26 October 2012, pp. 15–20 (2012)
51. Soille, P.: On genuine connectivity relations based on logical predicates. In: Proceedings of 14th International Conference Image Analysis Processing, Modena, Italy, pp. 487–492 (2007)
52. Mehta, M., Rissanen, J., Agrawal, R.: Mdl-based decision tree pruning. In: KDD (1995)
53. Niblett, T., Bratko, I.: Learning decision rules in noisy domains. In: Proceedings of Expert Systems, Brighton, UK (1986)
54. Quinlan, J.: Simplifying decision trees. Int. J. Man-Mach. Stud. **27**(3), 221–234 (1987)

55. Gueguen, L., Datcu, M.: A similarity metric for retrieval of compressed objects: application for mining satellite image time series. IEEE Trans. Knowl. Data Eng. **20**(4), 562–575 (2008)
56. Linde, Y., Buzo, A., Gray, R.: An algorithm for vector quantizer design. IEEE Trans. Commun. **28**(1), 84–95 (1980)
57. Li, M., Vitanyi, P.: An Introduction to Kolmogorov Complexity and its Application. Springer (1997)
58. Sculley, D., Brodley, C.E.: Compression and machine learning: a new perspective on feature space vectors. In: Proceedings of Data Compression Conference, 2006. DCC 2006, pp. 332–341 (2006)
59. Gower, J., Ross, G.: Minimum spanning trees and single linkage cluster analysis. J. R. Stat. Soc. - Appl. Stat. **18**(1), 54–64 (1969)
60. Yianilos, P.N.: Data structures and algorithms for nearest neighbor search in general metric spaces. In: Proceedings of 4th Annual ACM-SIAM Symposium on Discrete Algorithms, SODA '93. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, pp. 311–321 (1993)
61. Ciaccia, P., Patella, M., Zezula, P.: M-tree: an efficient access method for similarity search in metric spaces. In: Jarke, M., Carey, M.J., Dittrich, K.R., Lochovsky, F.H., Loucopoulos, P., Jeusfeld, M.A. (eds.) Proceedings of 23rd International Conference Very Large Data Bases (VLDB), Athens, Greece, 25–29 August 1997, pp. 426–435. Morgan Kaufmann
62. Daschiel, H., Datcu, M.: Cluster structure evaluation of dyadic k-means for mining large image archives. Proc. SPIE Image Signal Process. Remote Sens. VIII **4885**(1), 120–130 (2003)
63. Bentley, J.: Multidimensional binary search trees used for associative searching. Commun. ACM **18**(9), 509–517 (1975)
64. Burges, C.J.C.: A tutorial on support vector machines for pattern recognition. Data Min. Knowl. Discov. **2**, 121–167 (1998)
65. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The weka data mining software: an update. SIGKDD Explor. **11**(1), 10–18 (2009)
66. Pacifici, F., Du, Q.: Foreword to the special issue on optical multiangular data exploitation and outcome of the 2011 grss data fusion contest. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. **5**(1), 3–7 (2012)
67. Bovolo, F., Bruzzone, L., King, R.: Introduction to the special issue on analysis of multitemporal remote sensing data. IEEE Trans. Geosci. Remote Sens. **51**(4), 1867–1869 (2013)
68. Klaric, M., Claywell, B., Scott, G., Hudson, N., Sjahputera, O., Li, Y., Barratt, S., Keller, J., Davis, C.: GeoCDX: an automated change detection and exploitation system for high-resolution satellite imagery. IEEE Trans. Geosci. Remote Sens. **51**(4), 2067–2086 (2013)
69. Heas, P., Datcu, M.: Modeling trajectory of dynamic clusters in image time-series for spatio-temporal reasoning. IEEE Trans. Geosci. Remote Sens. **43**(7), 1635–1647 (2005)
70. Petitjean, F., Inglada, J., Gancarski, P.: Satellite image time series analysis under time warping. IEEE Trans. Geosci. Remote Sens. **50**(8), 3081–3095 (2012)
71. Văduva, C., Costăchioiu, T., Pătraşcu, C., Gavăt, I., Lăzărescu, V., Datcu, M.: A latent analysis of earth surface dynamic evolution using change map time series. IEEE Trans. Geosci. Remote Sens. **51**(4), 2105–2118 (2013)
72. Bruzzone, L., Prieto, D.: Unsupervised retraining of a maximum likelihood classifier for the analysis of multitemporal remote sensing images. IEEE Trans. Geosci. Remote Sens. **39**(2), 456–460 (2001)
73. Rajan, S., Ghosh, J., Crawford, M.: Exploiting class hierarchies for knowledge transfer in hyperspectral data. IEEE Trans. Geosci. Remote Sens. **44**(11), 3408–3417 (2006)
74. Bruzzone, L., Marconcini, M.: Toward the automatic updating of land-cover maps by a domain-adaptation SVM classifier and a circular validation strategy. IEEE Trans. Geosci. Remote Sens. **47**(4), 1108–1122 (2009)
75. Tuia, D., Ratle, F., Pacifici, F., Kanevski, M., Emery, W.: Active learning methods for remote sensing image classification. IEEE Trans. Geosci. Remote Sens. **47**(7), 2218–2232 (2009)
76. Tuia, D., Pasolli, E., Emery, W.: Using active learning to adapt remote sensing image classifiers. Remote Sens. Env. **115**(9), 2232–2242 (2011)

77. Matasci, G., Tuia, D., Kanevski, M.: Svm-based boosting of active learning strategies for efficient domain adaptation. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. **5**(5), 1335–1343 (2012)
78. Persello, C., Bruzzone, L.: Active learning for domain adaptation in the supervised classification of remote sensing images. IEEE Trans. Geosci. Remote Sens. **50**(11), 4468–4483 (2012)
79. Tuia, D., Muñoz Marí, J., Gómez-Chova, L., Malo, J.: Graph matching for adaptation in remote sensing. IEEE Trans. Geosci. Remote Sens. **51**(1), 329–341 (2013)
80. Leiva-Murillo, J., Gomez-Chova, L., Camps-Valls, G.: Multitask remote sensing data classification. IEEE Trans. Geosci. Remote Sens. **51**(1), 151–161 (2013)
81. Baraldi, A.: Impact of radiometric calibration and specifications of spaceborne optical imaging sensors on the development of operational automatic remote sensing image understanding systems. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. **2**(2), 104–134 (2009)
82. Wilkinson, G.: Results and implications of a study of fifteen years of satellite image classification experiments. IEEE Trans. Geosci. Remote Sens. **43**(3), 433–440 (2005)
83. Pacifici, F.: An automatic atmospheric compensation algorithm for very high spatial resolution imagery. In: Society of Photo-optical Instrumentation Engineers Defense, Security, and Sensing (SPIE), 2012, April 2012
84. Pacifici, F.: An automatic atmospheric compensation algorithm for very high spatial resolution imagery and its comparison to FLAASH and QUAC. In: Joint Agency Commercial Imagery Evaluation (JACIE), April 2013
85. Roy, D.P., Borak, J.S., Devadiga, S., Wolfe, R.E., Zheng, M., Descloitres, J.: The modis land product quality assessment approach. Remote Sens. Env. **83**(1–2), 62–76 (2002)
86. Updike, T., Comp, C.: Radiometric Use of WorldView-2 Imagery. Technical report, DigitalGlobe, Inc. (2010)
87. Chander, G., Markham, B.L., Helder, D.L.: Summary of current radiometric calibration coefficients for Landsat MSS, TM, ETM+, and EO-1 ALI sensors. Remote Sens. Env. **113**(5), 893–903 (2009)
88. Frank, T., Tweddale, S., Knapp, D.: Variability of at-satellite surface reflectance from landsat tm and noaa avhrr in death valley national monument. Photogramm. Eng. Remote Sens. **60**(10), 1259–1266 (1994)
89. Schott, J.: Remote Sensing: The Image Chain Approach: The Image Chain Approach. Remote Sensing and Geographic Information Systems. Oxford University Press (2007)
90. Longbotham, N., Chaapel, C., Bleiler, L., Padwick, C., Emery, W., Pacifici, F.: Very high resolution multiangle urban classification analysis. IEEE Trans. Geosci. Remote Sens. **50**(4), 1155–1170 (2012)
91. Clark, R., Swayze, G., Wise, R., Livo, E., Hoefen, T., Kokaly, R., Sutley, S.: USGS digital spectral library splib06a: U.S. Geological Survey, Digital Data Series 231 (2007)
92. Schaepman-Strub, G., Schaepman, M.E., Painter, T.H., Dangel, S., Martonchik, J.V.: Reflectance quantities in optical remote sensing - definitions and case studies. Remote Sens. Env. **103**, 27–42 (2006)
93. Rahman, H., Pinty, B., Verstraete, M.M.: Coupled surface-atmosphere reflectance (CSAR) Model: 2. Semi-empirical surface model usable with NOAA advanced very high resolution radiometer data. J. Geophys. Res. **98**(D11), 20791–20801 (1993)
94. Pinty, B., Widlowski, J.L., Gobron, N., Verstraete, M., Diner, D.: Uniqueness of multiangular measurements. I. An indicator of subpixel surface heterogeneity from MISR. IEEE Trans. Geosci. Remote Sens. **40**(7), 1560–1573 (2002)
95. Su, L., Huang, Y., Chopping, M.J., Rango, A., Martonchik, J.V.: An empirical study on the utility of BRDF model parameters and topographic parameters for mapping vegetation in a semi-arid region with MISR imagery. Int. J. Remote Sens. **30**(13), 3463–3483 (2009)
96. Lucht, W., Schaaf, C., Strahler, A.: An algorithm for the retrieval of albedo from space using semiempirical BRDF models. IEEE Trans. Geosci. Remote Sens. **38**(2), 977–998 (2000)
97. Kimes, D., Harrison, P., Harrison, P.: Learning class descriptions from a data base of spectral reflectance with multiple view angles. IEEE Trans. Geosci. Remote Sens. **30**(2), 315–325 (1992)

98. Sandmeier, S., Deering, D.: Structure analysis and classification of boreal forests using airborne hyperspectral BRDF data from ASAS. Remote Sens. Env. **69**(3), 281–295 (1999)
99. Chopping, M., Rango, A., Ritchie, J.: Improved semi-arid community type differentiation with the NOAA AVHRR via exploitation of the directional signal. IEEE Trans. Geosci. Remote Sens. **40**(5), 1132–1149 (2002)
100. Armston, J., Scarth, P., Phinn, S., Danaher, T.: Analysis of multi-date MISR measurements for forest and woodland communities, Queensland, Australia. Remote Sens. Env. **107**(1–2), 287–298 (2007)
101. Verrelst, J., Clevers, J.G.P.W., Schaepman, M.E.: Merging the Minnaert-k parameter with spectral unmixing to map forest heterogeneity with CHRIS/PROBA data. IEEE Trans. Geosci. Remote Sens. **48**(11), 4014–4022 (2010)
102. Laurent, V.C.E., Verhoef, W., Clevers, J.G.P.W., Schaepman, M.E.: Inversion of a coupled canopy - atmosphere model using multi-angular top-of-atmosphere radiance data: a forest case study. Remote Sens. Env. **115**, 2603–2612 (2011)
103. Koukal, T., Atzberger, C.: Potential of multi-angular data derived from a digital aerial frame camera for forest classification. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. **5**(1), 30–43 (2012)
104. Knobelspiesse, K., Cairns, B., Schmid, B., Román, M., Schaaf, C.: Surface BRDF estimation from an aircraft compared to MODIS and ground estimates at the Southern Great Plains site. J. Geophys. Res. Atmos. **113**(D20105) (2008)
105. Shell, J.: Polarimetric Remote Sensing in the Visible to Near Infrared. Ph.D. thesis, Rochester Institute of Technology (2005)
106. Wanner, W., Li, X., Strahler, A.: On the derivation of kernels for kernel-driven models of bidirectional reflectance. J. Geophys. Res. Atmos. **100**(D10), 21077–21089 (1995)
107. Hill, M., Averill, C., Jiao, Z., Schaaf, C., Armston, J.: Relationship of MISR RPV parameters and MODIS BRDF shape indicators to surface vegetation patterns in an Australian tropical savanna. Can. J. Remote Sens. **34**(2), 247–267 (2008)
108. Lucht, W., Lewis, P.: Theoretical noise sensitivity of brdf and albedo retrieval from the eos-modis and misr sensors with respect to angular sampling. Int. J. Remote Sens. **21**(1), 81–98 (2000)
109. Lucht, W.: Expected retrieval accuracies of bidirectional reflectance and albedo from EOS-MODIS and MISR angular sampling. J. Geophys. Res. Atmos. **103**(D8), 8763–8778 (1998)
110. Bovolo, F., Bruzzone, L.: A theoretical framework for unsupervised change detection based on change vector analysis in the polar domain. IEEE Trans. Geosci. Remote Sens. **45**(1), 218–236 (2007)
111. Bovolo, F.: A multilevel parcel-based approach to change detection in very high resolution multitemporal images. IEEE Geosci. Remote Sens. **6**(1), 33–37 (2009)
112. Carvalho, J.O.A., Guimarães, R.F., Gillespie, A.R., Silva, N.C., Gomes, R.A.T.: A new approach to change vector analysis using distance and similarity measures. Remote Sens. **3**(11), 2473–2493 (2011)
113. He, C., Zhao, Y., Tian, J., Shi, P., Huang, Q.: Improving change vector analysis by cross-correlogram spectral matching for accurate detection of land-cover conversion. Int. J. Remote Sens. **34**(4), 1127–1145 (2013)
114. Radke, R., Andra, S., Al-Kofahi, O., Roysam, B.: Image change detection algorithms: a systematic survey. IEEE Trans. Image Process. **14**(3), 294–307 (2005)
115. Jensen, J.: Introductory Digital Image Processing: A Remote Sensing Perspective. Prentice Hall Series in Geographic Information Science. Prentice Hall (1996)
116. Niemeyer, I., Canty, M., Klaus, D.: Unsupervised change detection techniques using multispectral satellite images. In: Proceedings of the IEEE 1999 International Geoscience and Remote Sensing Symposium, 1999. IGARSS '99, vol. 1, pp. 327–329 (1999)
117. Gong, P.: Change detection using principal components analysis and fuzzy set theory. Can. J. Remote Sens. **19**, January 1993

118. Koetz, B., Widlowski, J.L., Morsdorf, F., Verrelst, J., Schaepman, M.: Suitability of the parametric model RPV to assess canopy structure and heterogeneity from multi-angular CHRIS-PROBA data. In: Proceedings of the 4th CHRIS/PROBA Workshop, 19–21 September 2006, vol. 1 (2006)
119. Widlowski, J.L., Pinty, B., Gobron, N., Verstraete, M., Diner, D., Davis, A.: Canopy structure parameters derived from multi-angular remote sensing data for terrestrial carbon studies. Clim. Change **67**, 403–415 (2004)

Chapter 4
Very-High-Resolution and Interferometric SAR: Markovian and Patch-Based Non-local Mathematical Models

Charles-Alban Deledalle, Loïc Denis, Giampaolo Ferraioli, Vito Pascazio, Gilda Schirinzi and Florence Tupin

Abstract This chapter is dedicated to very-high-resolution (VHR) SAR imagery, including interferometric applications. First, the principles of SAR data acquisition are presented as well as the different types of configurations. The widely adopted Gaussian complex model of fully developed speckle is described as well as more advanced statistical models for VHR SAR data that account for textures. The following two parts are devoted to SAR image estimation and to image denoising within two different frameworks. First, Markovian modeling is introduced and the associated optimization approaches are presented, including graph-cut-based optimization. The second framework is the patch-based non-local modeling of SAR complex data. Both frameworks are adapted to SAR images through the use of statistical models specific to SAR imagery. Their applications to amplitude data, interferometry, and fusion with optical data are illustrated. A special focus is given to phase unwrapping

C.-A. Deledalle
IMB, CNRS, University of Bordeaux, Bordeaux INP, 33405 Talence, France
e-mail: charles-alban.deledalle@math.u-bordeaux.fr

L. Denis
Univ Lyon, UJM-Saint-Etienne, CNRS, Institut d Optique Graduate School,
Laboratoire Hubert Curien UMR 5516, F-42023 Saint-Etienne, France
e-mail: loic.denis@univ-st-etienne.fr

G. Ferraioli
Dipartimento di Scienze e Tecnologie, Università degli Studi di Napoli Parthenope,
Centro Direzionale di Napoli, Is. C4, 80143 Naples, Italy
e-mail: giampaolo.ferraioli@uniparthenope.it

V. Pascazio · G. Schirinzi
Dipartimento di Ingegneria, Università degli Studi di Napoli Parthenope,
Centro Direzionale di Napoli, Is. C4, 80143 Naples, Italy
e-mail: vito.pascazio@uniparthenope.it

G. Schirinzi
e-mail: gilda.schirinzi@uniparthenope.it

F. Tupin (✉)
LTCI, Telecom ParisTech, Université Paris Saclay, 46 rue Barrault, 75013 Paris, France
e-mail: florence.tupin@telecom-paristech.fr

G. Moser and J. Zerubia (eds.), *Mathematical Models for Remote Sensing Image Processing*, Signals and Communication Technology,
https://doi.org/10.1007/978-3-319-66330-2_4

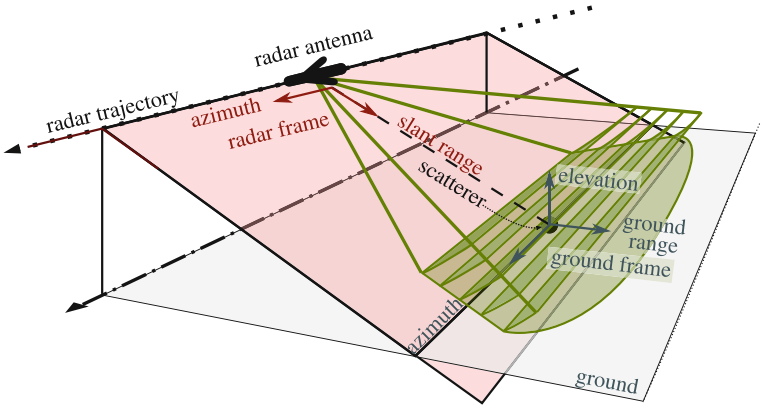


Fig. 4.1 Side-looking radar imagery: different time delays define different slices in 3D. Two frames are used: the radar frame (azimuth and slant range) and the ground frame (azimuth, ground range, and elevation)

applied to single- and multi-channel interferometry, showing the usefulness of local and global contextual models.

4.1 Principles of SAR Imagery

4.1.1 Principles of SAR Acquisition

Synthetic aperture radar (SAR) is an active system that emits microwave radiations toward the ground and measures the electromagnetic field backscattered by the illuminated area[1]. By means of a specific signal processing chain, the received signal is transformed into a high-resolution image of the observed scene.

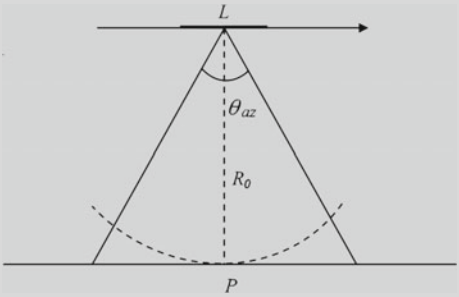
SAR, like real aperture radar (RAR) systems, is able to produce images from the backscattered signal, starting from the measurement of time delays between transmitted and received signals. The time delay is directly proportional to the distance between the sensor and the scatterer.

The obtained images are available in the two radar conventional coordinates, *azimuth* and *range*. The azimuth is the flight direction, while the range, which is orthogonal to the first one, is the looking direction of the antenna (i.e., the direction from the sensor to the scatterer). Radar acquisition geometry is shown in Fig. 4.1.

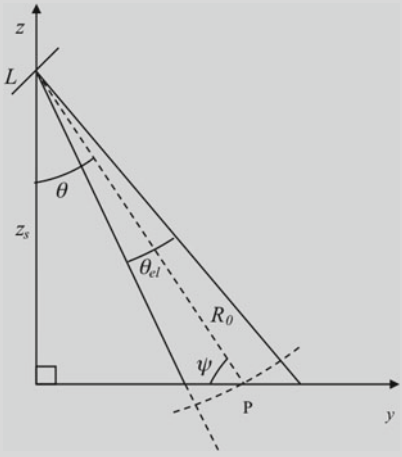
[1]The polarimetric case is mentioned in this chapter for completeness and to stress the generality of the described methods; it will be discussed more in-depth in the following chapter.

Real Aperture Radar Resolution

The acquisition geometry of RAR in the azimuth range plane is depicted in the following figure.



The acquisition geometry can also be shown in the elevation-ground range plane. The definition of the former is straightforward, while the latter is the direction orthogonal to the azimuth on the ground. The acquisition geometry in elevation-ground range plane is shown in the following figure.



RAR and SAR systems have different resolutions along the two directions (azimuth and range). The resolution is defined as the minimum distance between two scatterer points to be resolved (i.e., separated).

In the range direction, to achieve a higher resolution, a linear frequency-modulated signal is transmitted, the *chirp*. Using a chirp, the energy of the signal is spread over a larger bandwidth W. The range resolution is then a function of the bandwidth (W) of the transmitted signal and the achievable range resolution is given by [1]:

$$\text{res}_r = \frac{c}{2W} \tag{4.1}$$

where c is the speed of light. The range resolution is independent to the distance between the sensor and the scatterer points.

The given definition of the range is the one referred to as *slant range*. In terms of spatial resolution in a ground frame, another definition of range is used: the *ground range*. The relation between these two resolution definitions is given by:

$$\text{res}_g \simeq \frac{\text{res}_r}{\cos\psi} \tag{4.2}$$

where ψ is the so-called grazing angle, which is defined as the angle between the radar line of sight and the horizontal plane at the point of the reflection on the Earth (with plane Earth assumption). The ground range resolution is coarser than the slant range resolution and varies along the line of sight.

The azimuth resolution at a given range R_0 is given by [2]:

$$\text{res}_a \propto \frac{\lambda R_0}{L} \tag{4.3}$$

where λ is the wavelength and L is the antenna length. To increase the azimuth resolution, it is thus necessary to use larger antennas or work at a smaller distance from the scatterers. Another solution is to use a *synthetic* aperture as done by the SAR systems.

The substantial difference between SAR and RAR is the largely improved azimuth spatial resolution of SAR systems. This better resolution is achieved by synthesizing a larger antenna a posteriori while using a small antenna during radar measurements. The system takes advantage of the fact that the response of a point scatterer (target) on the ground is contained in more than a single radar echo and shows a typical *history of phase* during the observation time. A scatterer point, in fact, remains inside the antenna beam for a significant period of time and is observed by the SAR from different positions during the movement of the antenna along the trajectory. By coherently combining the different echos relative to the target, SAR realizes a *synthetically enlarged antenna*, a synthetic array able to increase the azimuth resolution.

Let us now focus on the processing chain for the image formation starting from the acquired data. Under the assumption that the electromagnetic interaction between the incident wave and the observed surface is dominated by surface scattering, the SAR system can be modeled as a linear model [3]. Let us consider the acquisition system of Fig. 4.2. We assume that a single point scatterer (target) is located on the ground with coordinates (x, r, θ), for the azimuth, slant range and elevation, respectively, at

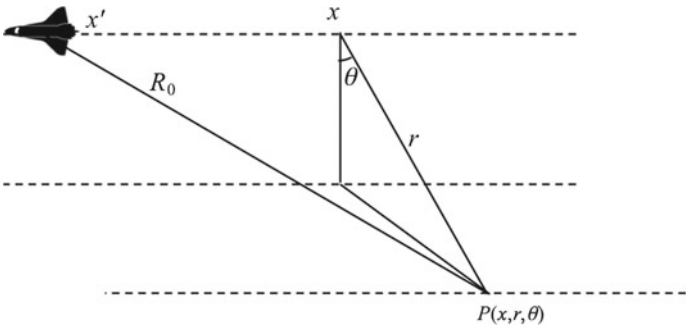


Fig. 4.2 Geometry of SAR Systems

a distance R_0 from the sensor. The sensor is moving in the azimuth direction with a constant velocity.

Let us establish how the target is seen by the SAR sensor. Assuming that a chirp modulation is adopted, the transmitted signal pulse at times $t_n - (\tau/2)$ is given by [4]:

$$d_{tx}(t - t_n) = \exp\left[j2\pi f\,(t - t_n) - j\frac{a}{2}\,(t - t_n)^2\right] \text{rect}\left(\frac{t - t_n}{\tau}\right) \tag{4.4}$$

where a is the chirp rate, f is the working frequency, τ is the duration of the pulses, and rect() is the rectangular function. In Eq. (4.4), all the amplitude factors have been neglected. The signal backscattered from the target and received at the sensor position (neglecting the backscattering coefficient) is given by:

$$d_{rx}(x' - x, t - t_n, r) = d_{tx}\left(t - t_n - \frac{2R_0}{c}\right) F\left(x' - x, r\right) \tag{4.5}$$

where F is a function related to the gain pattern of the transmit antenna.
Changing the notations ($t' = t - t_n$), moving from time to space ($r' = \frac{ct'}{2}$), introducing the wavelength λ, defining $\Delta R_0 = R_0 - r$, and after suppressing the fast varying term, we can define the unit response function g as:

$$g(x' - x, r - r', r) = d_{rx}(x' - x, r - r', r) \exp\left(j\frac{4\pi}{\lambda}r\right) =$$
$$\exp\left[-j\frac{4\pi}{\lambda}\Delta R_0\right] \exp\left[-j\frac{a}{2}\left(\frac{2}{c}\left(r' - \Delta R_0 - r\right)\right)^2\right] \text{rect}\left(\frac{r' - \Delta R_0 - r}{\frac{\tau c}{2}}\right) F\left(x' - x, r\right) \tag{4.6}$$

Considering the backscattering coefficient $\gamma(x, r)$ of the entire imaged area, the raw received signal is given by the superimposition of all the elementary contributions coming from the scene:

$$h(x', r') = \iint \gamma(x, r) \exp\left(-j\frac{4\pi}{\lambda}r\right) g(x' - x, r' - r, r) dx dr = \iint \bar{\gamma}(x, r) g(x' - x, r' - r, r) dx dr \tag{4.7}$$

where $\bar{\gamma}(x, r) = \gamma(x, r) \exp\left(-j\frac{4\pi}{\lambda}r\right)$. Moving to the frequency domain, the Fourier transform (FT) of the raw signal is given by:

$$H(\zeta, \eta) = \iint h(x', r') \exp(-jx'\zeta) \exp(-jr'\eta) dx' dr' \tag{4.8}$$

where ζ and η are the Fourier domain variables corresponding to x' and r', respectively. By using Eq. (4.7) and introducing $G()$ as the FT of $g()$, it is possible to rewrite Eq. (4.8) as:

$$H(\zeta, \eta) = \iint \bar{\gamma}(x, r) G(\zeta, \eta, r) \exp(-jx\zeta) \exp(-jr\eta) dx dr \tag{4.9}$$

The SAR imaging problem consists in designing an appropriate filter able to recover an optimal estimation of $\bar{\gamma}$ from Eq. (4.9). The filter has to be able to properly filter out the function G. To this aim, the analytical evaluation of the function is needed [5]. The main problem to be faced out with the function G is related to its explicit dependence on the range variable r.

Synthetic Aperture Radar Resolution

Let us focus on the case of a single point scatterer with coordinates (x_0, r_0). The backscattering coefficient becomes $\bar{\gamma}(x, r) = \bar{\gamma}(x_0, r_0)\delta(x - x_0)\delta(r - r_0)$. Thus, exploiting the property of Dirac function δ, the raw received signal can be written as:

$$h(x', r') = \bar{\gamma}(x_0, r_0) g(x' - x_0, r' - r_0, r_0) \tag{4.10}$$

and its FT:

$$H(\zeta, \eta) = \iint \bar{\gamma}(x_0, r_0) g(x' - x_0, r' - r_0, r_0) \exp(-jx'\zeta) \exp(-jr'\eta) dx' dr' = \bar{\gamma}(x_0, r_0) G(\zeta, \eta, r_0) \exp(-jx_0\zeta) \exp(-jr_0\eta) \tag{4.11}$$

Multiplying Eq. (4.11) by conjugate function $G^*(\zeta, \eta, r_0)$ and by taking the inverse FT of the result, it is possible to estimate the focused image.

$$\hat{\gamma}(x', r') = \iint H(\zeta, \eta) G^*(\zeta, \eta, r_0) \exp(jx'\zeta) \exp(jr'\eta) d\zeta d\eta \tag{4.12}$$

and exploiting the stationary phase method [5] for the approximation of the $G()$ function:

$$\begin{aligned} &\hat{\gamma}(x', r') = \\ &= \iint \bar{\gamma}(x_0, r_0) \operatorname{rect}\left(\frac{\zeta}{2e}\right) \operatorname{rect}\left(\frac{\eta}{2b}\right) \exp(j(x' - x_0)\zeta) \exp(j(r' - r_0)\eta) d\zeta d\eta = \\ &\qquad \bar{\gamma}(x_0, r_0) \operatorname{sinc}\left((x' - x_0)e\right) \operatorname{sinc}\left((r' - r_0)b\right) \end{aligned} \tag{4.13}$$

where $e = \frac{\pi X}{L/2}$ and $b = a\tau^2/2$, with X being the length of the synthetic antenna and the others parameters previously introduced. Further details on these definitions and on the approximation of the $G()$ function can be found in [3].

The result of Eq. (4.13) can be used in order to define the spatial resolution, in both range and azimuth coordinates, of the SAR systems compared to RAR ones. Following the approach of [3], the azimuthal and range normalized (with respect to the RAR resolution) achievable resolutions are given by the distance between the -3 decibel points of the two $sinc$ functions:

$$\text{res}_{a_{norm}} = \frac{2\pi}{2e} = \frac{L^2}{2\lambda R_0} \tag{4.14}$$

$$\text{res}_{r_{norm}} = \frac{2\pi}{2b} = \frac{2\pi}{a\tau^2} \tag{4.15}$$

It is easy to verify that the two previous values are much smaller than the unit, using typical SAR parameters. Moreover, by multiplying Eqs. (4.14) and (4.15) by Eqs. (4.3) and (4.1), respectively, it is possible to obtain the SAR achievable resolutions.

According to the previously explained formulation, a properly designed $G^*()$ function, calculated for each range coordinate, is mandatory to estimate the backscattering signal. Its dependence to the range coordinate makes the processing complex and not straightforward. Several solutions have been adopted in literature [5]. In the following, we report one of them which is known as *grid rectification* approach. For this aim, the $G(\zeta, \eta, r)$ is factorized as the product of two functions:

$$G(\zeta, \eta, r) = G_0(\zeta, \eta) \Lambda(\zeta, \eta, r) \tag{4.16}$$

where only the second function depends on r. Using some approximations [5], the function can be conveniently expressed as:

$$\Lambda(\zeta, \eta, r) = \exp\left(j\frac{\lambda\zeta^2}{8\pi}r\right) \tag{4.17}$$

The FT of the raw signal of Eq. (4.9) after substituting Eqs. (4.16) and (4.17) becomes:

$$H(\zeta, \eta) = \iint \bar{\gamma}(x, r)G_0(\zeta, \eta)\Lambda(\zeta, \eta, r)\exp(-jx\zeta)\exp(-jr\eta)dxdr \tag{4.18}$$

Making a change of variables $\eta_c = \eta + \frac{\lambda\zeta^2}{8\pi}r$, the previous equation can be written as:

$$H(\zeta, \eta) = G_0(\zeta, \eta)\bar{\Gamma}(\zeta, \eta_c) \tag{4.19}$$

where $\bar{\Gamma}()$ is the FT of $\bar{\gamma}()$. Finally, by multiplying for the conjugate function ${G_0}^*(\zeta, \eta)$, it is possible to estimate the FT of the focused image:

$$\hat{\Gamma}(\zeta, \eta) = H(\zeta, \eta){G_0}^*(\zeta, \eta) = \bar{\Gamma}(\zeta, \eta_c) \tag{4.20}$$

It is clear that the estimated function computed in the uniform rectangular grid (ζ, η) corresponds to a function computed over a grid (ζ, η_c) which is parabolic in one direction (there is a dependence on the square of ζ). Thus, a final rectification of the parabolic grid onto the rectangular one is mandatory. Details on this rectification can be found in [5].

4.1.2 From 2D to 4D SAR Imaging

The term SAR interferometry (InSAR) refers to all methods that employ at least two complex SAR images to derive more information about an object compared to the information provided by a single SAR image [6]. Two possible configurations of SAR interferometry exist: *across track interferometry* and *along track interferometry*. In the across track configuration, two SAR sensors fly on two parallel flight lines and look at the ground from slightly different positions. Across track interferometry is used to reconstruct Earth topography, providing high-precision digital elevation model (DEM) of Earth surface. In the along track configuration, the two sensors fly on the same flight direction, looking the scene from the same position but with a small temporal delay. This system is mainly used for measurement of ocean currents [7] and for moving target detection [8]. In the following, the first configuration will be analyzed.

The geometry of an interferometric SAR system is shown in Fig. 4.3. Two SAR systems (SAR1 and SAR2) observe the scene from two different positions, acquiring two complex images. From the analysis reported in the previous section, the two focused, $\hat{\gamma}_1(x', r')$ and $\hat{\gamma}_2(x', r')$, differ for a phase term proportional to dR, which

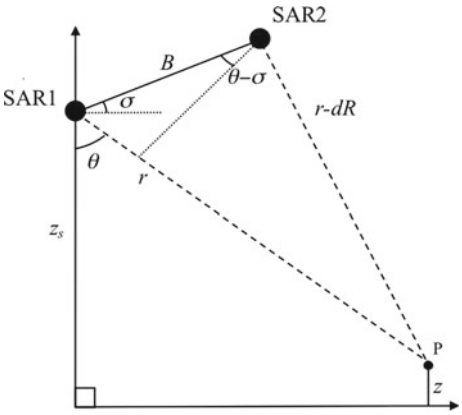


Fig. 4.3 Interferometric SAR geometry. The point P is observed from SAR1 and SAR2 from two different positions. The distance between the sensors B is called the baseline. The height of the observed point is inferred from the measure of the additional range path dR

is the additional range path of SAR2 with respect to the distance between the target and SAR1.

By taking the angle of the conjugate product of the two complex images, it is possible to define the so-called interferometric phase:

$$\phi(x', r') = \arg\left[\hat{\gamma}_1(x', r')\hat{\gamma}_2^*(x', r')\right] = -\frac{4\pi}{\lambda} dR \tag{4.21}$$

From Eq. (4.21), it is clear that the phase difference between the two images (the interferogram) provides an accurate measure of the difference in range [2, 6, 9]. The additional range path $-dR$ can be estimated with high precision.

Starting from the knowledge of interferometric phase ϕ, it is possible to infer the height of the observed scene. Let us consider the geometry depicted in Fig. 4.3.

The distance B between the two sensors is the so-called *baseline*, the angle between the line of sight and the vertical direction θ is the so-called *look angle*, while σ is the angle between the horizontal direction and the baseline. Using some geometrical relations and some approximations, it is easy to show that dR is related to the baseline by the following relation:

$$dR = -B \sin(\theta - \sigma) \tag{4.22}$$

while the height of the point can be computed from the evaluation of the look angle θ using the relation:

$$z = z_s - r\cos(\theta) \tag{4.23}$$

where z_s is the height of the sensor. It is interesting to evaluate the relationship between a change in the scatterer height Δz and the resulting change in the difference in range to the two phase receivers ΔdR:

$$\frac{\Delta z}{\Delta dR} = \frac{r \sin(\theta)}{B\ cos\ (\theta - \sigma)} \tag{4.24}$$

Exploiting Eq. (4.21) and rearranging the terms, it is possible to evaluate the change in the interferometric phase due to a change in height:

$$\Delta\phi = \frac{4\pi B \cos(\theta - \sigma)}{\lambda r \sin\theta} \Delta z \tag{4.25}$$

This relation provides the interferometer height sensibility and provides the relation between the measured phase and the height in the absence of noise.

An evolution of SAR interferometry is the differential interferometry (DInSAR), which is a technique able to estimate not only the height of pixels but also their possible displacements, occurred between two successive observations. Considering a displacement of the pixel along the line of sight in two observations, it is possible to rewrite Eq. (4.21) in the following way:

$$\phi(x, r) = -\frac{4\pi}{\lambda} dR = -\frac{4\pi}{\lambda} d R_d - \phi_z \tag{4.26}$$

where ϕ_z is the contribution corresponding to the target height while $d R_d$ is the range variation corresponding to the displacement. The terms associated with the variation of the wave propagation delay through the atmosphere and the phase noise have been neglected [2]. DInSAR allows measuring the component of the displacement with an accuracy of the order of fractions of the wavelength. A deeper analysis and discussion on DInSAR can be found in [2].

4.1.3 Statistics of Speckle in SAR Imagery

4.1.3.1 Fully Developed Speckle Model

SAR data are represented by a matrix of complex numbers representing the backscattered electromagnetic field. Both amplitude and phase can be useful depending on the considered application. As we shown in the previous section, phase differences of interferometric images are related to the scene geometry and provide information on ground elevation or movement. On the other hand, amplitude (absolute value of the complex field) represents the backscattering properties of the illuminated surfaces. Besides, if different polarizations for wave emission and reception are used, a complete characterization of the physical properties of the surface is provided through the polarimetric scattering vector [10]. More details on polarimetric SAR can be found in Chap. 5.

Whatever the considered parameters (amplitude, phase, polarimetric scattering vector,...), they present high fluctuations due to the speckle phenomenon. It is linked to the coherent nature of radar illumination and to the interferences of elementary

scatterers inside the resolution cells. This phenomenon has been deeply studied and modeled by Goodman [11]. In the case of rough surfaces compared to the electromagnetic wavelength, it can be established that for a given surface of reflectivity R, the complex electromagnetic field[2] Z follows a circular Gaussian distribution with zero mean defined by:

$$\mathrm{p}(\mathrm{Z}|\mathrm{R}) = \mathrm{p}(\Re(\mathrm{Z}), \Im(\mathrm{Z})|\mathrm{R}) = \frac{1}{\pi \mathrm{R}} \exp\left(-\frac{|\mathrm{Z}|^2}{\mathrm{R}}\right) \tag{4.27}$$

4.1.3.2 Speckle Model for Amplitude or Intensity Data

This complex circular Gaussian distribution implies an exponential distribution for intensity ($\mathrm{I} = |\mathrm{Z}|^2$):

$$\mathrm{p}(\mathrm{I}|\mathrm{R}) = \frac{1}{\mathrm{R}} \exp\left(-\frac{\mathrm{I}}{\mathrm{R}}\right) \qquad I \geq 0 \tag{4.28}$$

a Rayleigh distribution for amplitude data ($\mathrm{A} = |\mathrm{Z}|$):

$$\mathrm{p}(\mathrm{A}|\mathrm{R}) = \frac{2\mathrm{A}}{\mathrm{R}} \exp\left(-\frac{\mathrm{A}^2}{\mathrm{R}}\right) \qquad \mathrm{A} \geq 0 \tag{4.29}$$

and a uniform phase distribution on the $[-\pi; \pi]$ interval.

To reduce speckle effects and the high variability of the signals, multi-look processing is a widespread technique finding its grounding in the maximum likelihood estimation. Multi-look intensity is obtained by averaging L incoherent intensity data. The distribution is then given by a Gamma probability density function (pdf):

$$\mathrm{p}_L(\mathrm{I}|\mathrm{R}) = \frac{L^L}{\mathrm{R}^L \Gamma(L)} \mathrm{I}^{(L-1)} \exp\left(-\frac{L\mathrm{I}}{\mathrm{R}}\right) \tag{4.30}$$

with Γ representing the Gamma function [12]. A usual modeling of SAR amplitude or intensity data is given by the multiplicative model separating the scene contribution and the speckle fluctuation effect. It is written $\mathrm{I} = \mathrm{R} \times \mathrm{S}$ where S represents the pure speckle following the previously given distributions with $\mathrm{R} = 1$, giving:

$$\mathrm{p}_L(\mathrm{S}) = \frac{L^L}{\Gamma(L)} \mathrm{S}^{(L-1)} \exp(-L\mathrm{S}) \tag{4.31}$$

[2]Corresponding to the backscattering coefficient $\gamma(x, r)$ of the previous section but taking into account the amplitude factors.

for multi-look intensity data. This formulation shows why Rayleigh or Gamma-distributed data have signal-dependent fluctuations (the variance depends on the underlying R value).

A well-known transform when dealing with SAR images is the *homomorphic* transform obtained when applying a logarithm to the data. This transform has many advantages as we will see in the following sections, since it converts the multiplicative speckle to an additive one, thus isolating the noise contribution. The transformed signal follows a Fisher–Tippett pdf [11] and has a stabilized variance $\sigma^2 = \psi^{(1)}(L)$ depending only on the number of looks L and no more on the scene value (with $\psi^{(1)}$ the polygamma function [12]).

4.1.3.3 Vectorial Speckle Model

If instead of considering a single complex value and associated quantities, a vector of complex values is considered (for instance, a polarimetric scattering vector or an interferometric vector), a zero mean complex circular Gaussian is followed:

$$\mathrm{p}(\mathbf{k}|\boldsymbol{\Sigma}) = \frac{1}{\pi^d \ \det(\boldsymbol{\Sigma})} \exp(-\mathbf{k}^{*t} \boldsymbol{\Sigma}^{-1} \mathbf{k}) \tag{4.32}$$

with $\boldsymbol{\Sigma}$ the covariance matrix determining the imaged surface.

Although limited to rough and homogenous surfaces, Goodman's model of fully developed speckle is widely used to process SAR data and estimate the physical parameters. It allows to predict the fluctuations of the parameters depending on the considered surface characterized by R (the reflectivity) or $\boldsymbol{\Sigma}$ for single- or multi-channel data, respectively.

In the same way as for intensity or amplitude, multi-looking is widely used to reduce signal fluctuations. In the vectorial case, multi-looking of multivariate Gaussian vectors of D dimension is done through the hermitian product, leading to the empirical covariance matrix:

$$\mathbf{C} = \frac{1}{L} \sum_{i=1}^{L} \mathbf{k}_i \mathbf{k}_i^{t*}$$

It follows the Wishart distribution when $L \geq D$:

$$\mathrm{p}(\mathbf{C}|\boldsymbol{\Sigma}) = \frac{L^{LD} |\mathbf{C}|^{L-D} \exp\left(-L\mathrm{Tr}(\boldsymbol{\Sigma}^{-1}\mathbf{C})\right)}{\pi^{\frac{D(D-1)}{2}} \Gamma(L)\Gamma(L-D+1) \ |\boldsymbol{\Sigma}|^L} \tag{4.33}$$

The samples to apply multi-looking are usually taken in the surrounding neighborhood, for instance, using a local window. A trade-off between the variance reduction (with a high L value) and the bias introduction (caused when using samples from a different underlying scene) is a critical issue that will be discussed in Sect. 4.3.

4.1.3.4 Speckle Model for InSAR Data

A special case of vectorial data is given by interferometric data $\mathbf{k} = (\mathrm{Z}_1\ \mathrm{Z}_2)^t$. The covariance matrix is then the following:

$$\mathbf{\Sigma} = \begin{bmatrix} \mathbb{E}(|\mathrm{Z}_1|^2) & \mathbb{E}(\mathrm{Z}_1\mathrm{Z}_2^*) \\ \mathbb{E}(\mathrm{Z}_1^*\mathrm{Z}_2) & \mathbb{E}(|\mathrm{Z}_2|^2) \end{bmatrix} \tag{4.34}$$

with $\mathbb{E}$ denoting the expectation.[3] A useful quantity is the complex correlation coefficient:

Negative log-likelihood in SAR imagery:

Under Goodman's model, the negative log-likelihood for a pixel value is given below for the different considered SAR modalities (up to an additive constant).

- **SAR Amplitude:** $\ell(\mathrm{A}|\mathrm{R}) = L\log\mathrm{R} + L\dfrac{\mathrm{A}^2}{\mathrm{R}}$.
- **SAR Intensity:** $\ell(\mathrm{I}|\mathrm{R}) = L\log\mathrm{R} + L\dfrac{\mathrm{I}}{\mathrm{R}}$.
- **SAR Covariance Matrix:** $\ell(\mathbf{C}|\mathbf{\Sigma}) = L\log|\mathbf{\Sigma}| + L\mathrm{Tr}(\mathbf{\Sigma}^{-1}\mathbf{C})$.

$$\rho_{12}^c = \frac{\mathbb{E}(\mathrm{Z}_1\mathrm{Z}_2^*)}{\sqrt{\mathbb{E}(|\mathrm{Z}_1|^2)\,\mathbb{E}(|\mathrm{Z}_2|^2)}} = \rho\, e^{j\varphi} \tag{4.35}$$

ρ being the coherence (scalar) denoting the correlation between the two complex signals and φ representing the true interferometric phase.

Different distributions linking observed data and true values can be given [10]. We only present some of them. The joint distribution of the observed empirical elements of $\mathbf{C}$ conditionally to the true values is expressed as:

$$\begin{aligned} \mathrm{p}(\mathrm{I}_1, \mathrm{I}_2, \phi|\mathrm{R}_1, \mathrm{R}_2, \rho, \varphi) = {} & \frac{1}{\pi^2\mathrm{R}_1\mathrm{R}_2(1-\rho^2)} \\ & \exp\left(-\frac{1}{1-\rho^2}\left(\frac{\mathrm{I}_1}{\mathrm{R}_1} + \frac{\mathrm{I}_2}{\mathrm{R}_2} - \frac{2\sqrt{\mathrm{I}_1\mathrm{I}_2}\rho\cos(\phi-\varphi)}{\sqrt{\mathrm{R}_1\mathrm{R}_2}}\right)\right) \end{aligned} \tag{4.36}$$

[3] $\mathbb{E}$ is used here for the expectation operator, differently from the other chapters, for clarity and to avoid ambiguities.

As seen before, in practice, some multi-looking is applied through the hermitian product of L complex values to compute $\mathbf{C}$. The following pdf can be deduced from the Gaussian circular model:

$$\mathrm{p}(\phi|\rho,\varphi,L) = \frac{(1-\rho^2)^L}{2\pi}\frac{1}{2L+1}\,{}_2F_1\left(2,2L;L+\frac{3}{2};\frac{1+\rho\cos(\phi-\varphi)}{2}\right) \tag{4.37}$$

where ${}_2F_1$ is a hypergeometric function [12]. In the case of 1-look data, the phase distribution can be expressed as [13]:

$$\mathrm{p}(\phi|\rho,\varphi) = \frac{1}{2\pi}\frac{1-\rho^2}{1-\rho^2\cos(\phi-\varphi)^2}\left(1+\frac{\rho\cos(\phi-\varphi)\cos^{-1}(-\rho\cos(\phi-\varphi))}{(1-\rho^2\cos(\phi-\varphi)^2)^{1/2}}\right) \tag{4.38}$$

4.1.3.5 Taking Texture into Account

The usual complex multivariate Gaussian model presented above is very efficient and depends only on a few parameters (reflectivity or covariance matrix). Nevertheless, to take into account texture information, more sophisticated distributions can be introduced which are derived from the multiplicative model. Introduced in [14], the Mellin transform dedicated to positive values, the Mellin convolution dealing with product of positive random variables, and the log-cumulant framework provide efficient tools to derive advanced models like the Generalized Gamma or Fisher distributions for amplitude data [15–17] or the Kummer-U for polarimetric data [18].

The Mellin framework will be discussed in more details in Chap. 5, especially with focus on the polarimetric case, but we will introduce here the diagram of log-cumulants of order 2 and 3 ($\widetilde{\kappa_2},\widetilde{\kappa_3}$). Denoting $\tilde{m}_k = \int_0^{+\infty}(\log u)^k p(u)du$, the log-cumulants are defined by:

$$\begin{aligned}\widetilde{\kappa_2} &= \tilde{m}_2 - \tilde{m}_1^2\\ \widetilde{\kappa_3} &= \tilde{m}_3 - 3\tilde{m}_2\tilde{m}_1 + 2\tilde{m}_1^3\end{aligned}$$

They can be simply computed through the empirical mean of logarithmic data.

This diagram (Fig. 4.4) provides a graphical representation of the different distributions that can be defined on $\mathbb{R}^+$ based on their log-cumulants of order 2 and 3 (i.e., not taking into account the reflectivity of the scene but the shape—head and tail—of the associated pdf).

These advanced models are useful when dealing with very-high-resolution data for which the texture has to be taken into account. A unifying distribution, the Meijer pdf, allows an almost full coverage of the log-cumulant diagram [19]. Nevertheless, the

modeling improvement is earned at the price of higher-order parameter estimation. In the following sections, we will focus on approaches exploiting the basic Goodman's model (Gamma and Wishart pdf) for the data but introducing textural information through the spatial correlation of neighboring pixels.

4.2 Markovian Modeling and Its Applications

4.2.1 Markovian Modeling of Images

4.2.1.1 The Markovian Framework

Section 4.1.3 described several statistical models to account for fluctuations due to speckle or scene heterogeneity. Starting from a SAR image, many tasks require to infer a spatial distribution of physical parameters (e.g., reflectivity, interferometric phase, polarimetric properties) or of higher-level attributes (class index for classification, indicator function for segmentation, change indicator for change detection). Beyond the statistical modeling of fluctuations in the data, it is beneficial to also capture the statistical dependency of the spatial distribution of interest. Indeed, when it comes to inferring a parameter or attribute at a given location in a SAR image, not only the observed SAR signal at that location should be used but also the spatial context.

Modeling the statistical dependence between random variables is a very old topic in science. The major difficulty here comes from the huge dimension of the random vector[4] $\mathbf{u}$ of parameters/attributes, each element u_i of the vector defining a different spatial location: from a few hundreds (one per image segment) to millions or billions (several parameters per pixel). To alleviate this complexity issue, the probability density function of the random vector[5] $\mathbf{u}$ is generally supposed to factor into a product of terms each involving only a small subset of elements of $\mathbf{u}$:

$$\mathrm{p}(u_1, \ldots, u_n) = \prod_k f_k\left(u_{I_1^k}, \ldots, u_{I_{n_k}^k}\right), \tag{4.39}$$

where indices I_1^k to $I_{n_k}^k$ define the kth subvector of dimension n_k. Such a factorization is often represented by a graph and each term in Eq. (4.39) is called a factor, see the box "Graphical models used to capture statistical dependence" below.

[4] In the following, we will denote by $\mathbf{u}$ the parameters of interest and $\mathbf{v}$ the observations provided by the data in a generic way. If useful, these notations will be replaced by the notations introduced in Sect. 4.1.3 for SAR data.

[5] Since each element of the random vector can be assigned to a given spatial location, we will refer to this random vector as a *random field*.

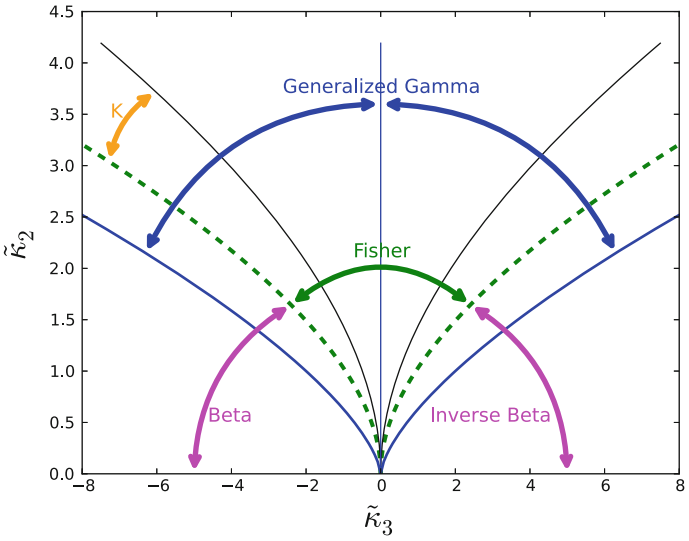


Fig. 4.4 Pdf positioning in the $\widetilde{\kappa_2} - \widetilde{\kappa_3}$ diagram. Each pdf is represented by a point in the diagram depending on its parameter. Pdf with one varying parameter (in addition to the mean value) covers a curve in this diagram while 2-parameter pdf covers an area. Generalized Gamma covers the area between blue curves excluding ordinate axis; Fisher, the area between dashed green curves; K pdf, the yellow area; and lognormal distributions stand along the ordinate axis, while beta and inverse beta are indicated in *pink areas* (color figure online)

In order to obtain a factorization of the form of Eq. (4.39), a (local) *Markov property* is typically assumed. A random field $\mathbf{u}$ is said to fulfill a (local) *Markov property* if and only if the conditional dependence is local:

$$\forall i,\ \mathrm{p}\big(u_i \,\big|\, u_1, \ldots, u_{i-1}, u_{i+1}, \ldots, u_n\big) = \mathrm{p}\big(u_i \,\big|\, \{u_k | k \in \mathscr{N}_{(i)},\ k \neq i\}\big) \tag{4.40}$$

where $\mathscr{N}_{(i)}$ is the set of indices of all the neighbors of site i. In order to define neighborhood relationships between sites, an undirected graph is associated with the Markov random field. Each node of the graph represents a different site (i.e., an element of the random vector, which also corresponds to a given spatial location). Two sites that are neighbors are connected by an edge in the graph; hence, the neighborhood $\mathscr{N}_{(i)}$ corresponds to all nodes that are connected to node i by an edge. Since the notion of neighborhood is central to the Markovian property, subgraphs composed of mutually connected nodes play an important role. By definition, *cliques* correspond to complete subgraphs. Clifford–Hammersley theorem [20] states (provided that no configuration of the random field is forbidden) that distributions that fulfill the Markov property are members of the family of *Gibbs distributions*, i.e., they can be written under the form:

$$\mathrm{p}(\mathbf{u}) = \frac{1}{\mathscr{Z}} \prod_{c \in \mathscr{C}} \psi_c(\mathbf{u}_c) = \frac{1}{\mathscr{Z}} \exp\Big\{ - \sum_{c \in \mathscr{C}} \theta_c(\mathbf{u}_c) \Big\} \tag{4.41}$$

where $\mathscr{Z}$ is a normalization constant (the *partition function*), c is a clique from the set $\mathscr{C}$ of all cliques of the graph, $\mathbf{u}_c$ is the subvector obtained by restricting vector $\mathbf{u}$ to the sites in the clique c, functions ψ_c are called *potential functions* while functions $\theta_c(\cdot) = -\log(\psi_c(\cdot))$ are the *clique energies*.[6]

It is sometimes desirable to design clique energies θ_c that can be driven by the data. For example, one may think of a clique energy that assigns a large value to configurations in which two neighboring sites i and j take very different parameter values u_i and u_i *except* if the observed values v_i and v_j are themselves very different. Conditional random fields (CRFs) have been introduced [21] so that clique energies can depend on the vector of observations. The probability density function of a CRF, conditionally to the observations $\mathbf{v}$, can be factored under a form similar to MRFs:

$$\mathrm{p}(\mathbf{u}|\mathbf{v}) = \frac{1}{\mathscr{Z}} \prod_{c \in \mathscr{C}} \psi_c(\mathbf{u}_c, \mathbf{v}) = \frac{1}{\mathscr{Z}} \exp\Big\{ - \sum_{c \in \mathscr{C}} \theta_c(\mathbf{u}_c, \mathbf{v}) \Big\} \tag{4.42}$$

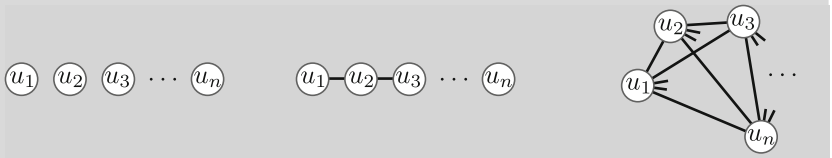
and the (local) Markov property is fulfilled, conditionally to the data:

$$\forall i,\ \mathrm{p}\big(u_i \,\big|\, \mathbf{v},\ u_1, \ldots, u_{i-1}, u_{i+1}, \ldots, u_n\big) = \mathrm{p}\big(u_i \,\big|\, \mathbf{v},\ \{u_k | k \in \mathscr{N}_{(i)},\ k \neq i\}\big) \tag{4.43}$$

We summarize here the three models discussed so far and illustrate their representation by graphs (*graphical models*):

Graphical models used to capture statistical dependence:

Between spatial independence (left) and complete dependence (right), graphical models describe conditional dependence of small subsets of the random variables:

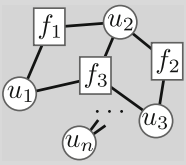


total independence a Markov random field complete dependence

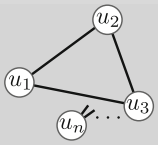
Three graphical models are widely used:

– *factor graphs* represent the factorization of the probability density function by connecting groups of variables through the factors f_k (see also Eq. (4.39))

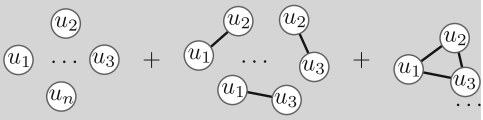
[6] The vocabulary in use in Markov random fields theory comes from the field of statistical physics.



– *Markov random fields* (MRFs) represent the factorization of the probability density function as a product over cliques (see also Eq. (4.41))

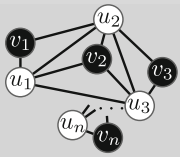


graphical model
of a Markov random field



decomposition into cliques

– *conditional random fields* (CRF) may include an external field (e.g., observations) to define the probability density function (see also Eq. (4.42))



graphical model
of a conditional random field

white circles: parameters of interest

black circles: observations

4.2.1.2 Applications in Remote Sensing

The Markovian framework is very popular for different tasks of image processing like denoising and regularization, classification, segmentation, or object extraction. Some of them will be described in details in the following sections. In the context of remote sensing applications, due to the high level of variability of SAR data, it has been widely used. The following references are only examples of such methods, but many other works have been developed.

The segmentation of SAR data with a Markov random field has been proposed in [22] with a dictionary of distributions and an automatic parameter learning, in [17] with Fisher distributions for VHR data and urban areas, in [23, 24] with Triplet Markov random field taking into account non-stationnarities, or hierarchical models [25].

This framework is well adapted whatever the processed data like polarimetric images [24, 26], or multi-temporal images [27, 28]. It has also been extensively studied for change detection applications [29].

Markovian framework is also very powerful for object extraction like roads [30] with Markov random fields defined on graphs of segments or connected components. An extension is provided by marked point processes [31], but these models are beyond the scope of this chapter. MRF models for multi-source and especially multi-resolution fusion will be discussed in Chap. 7.

In the next sections, applications of Markov random fields for amplitude and phase denoising with an auxiliary optic data will be described (Sect. 4.2.3), as well as phase unwrapping applications with multi-channel interferometry (Sect. 4.2.4).

4.2.2 *Inference in Markov Random Fields*

Based on the statistical model of the observations $\mathbf{v}$ (Sect. 4.1.3) and of the spatial distribution of parameters $\mathbf{u}$ (Sect. 4.2.1.1), application of Bayes rule leads to an expression of the posterior distribution of the form:

$$\mathrm{p}(\mathbf{u}|\mathbf{v}) \propto \prod_i \exp\{-\ell(v_i|u_i)\} \prod_{c\in\mathscr{C}} \exp\{-\theta_c(\mathbf{u}_c)\} \tag{4.44}$$

where we considered the observations v_i and v_j at two different sites i and j to be independent (conditionally to the parameters $\mathbf{u}$), and $\ell(v_i, u_i) = -\log \mathrm{p}(v_i|u_i)$ is the so-called neg-log-likelihood function for parameter u_i where $\mathrm{p}(v_i|u_i)$ is one of the pdfs introduced in Sect. 4.1.3.

Due to the coupling of all random variables u_i of the random vector $\mathbf{u}$ induced by the chaining of all cliques, estimation of $\mathbf{u}$ from the posterior distribution $\mathrm{p}(\mathbf{u}|\mathbf{v})$ is not easy. Two estimators are typically considered: the posterior mean estimator $\widehat{\mathbf{u}}^{(\mathrm{PM})} = \int \mathbf{u} \cdot \mathrm{p}(\mathbf{u}|\mathbf{v})\mathrm{d}\mathbf{u}$ and the maximum a posteriori $\widehat{\mathbf{u}}^{(\mathrm{MAP})} \in \arg\max_{\mathbf{u}} \mathrm{p}(\mathbf{u}|\mathbf{v})$. Computation of the posterior mean estimator requires to perform a high-dimensional integration, which is typically done using sampling methods like Monte Carlo Markov chains (MCMC) [32]. The computational complexity involved by these approaches and the lack of relevance of the mean when the posterior distribution is multimodal has led to favor the maximum a posteriori (MAP) estimation in many works. For posterior distributions of the form of Eq. (4.44), MAP estimation requires to solve an optimization problem:

$$\widehat{\mathbf{u}}^{(\mathrm{MAP})} \in \arg\min_{\mathbf{u}} \mathscr{F}_{\mathbf{v}}(\mathbf{u}), \quad \text{with } \mathscr{F}_{\mathbf{v}}(\mathbf{u}) = \sum_i \ell(v_i|u_i) + \sum_{c\in\mathscr{C}} \theta_c(\mathbf{u}_c) \tag{4.45}$$

We discuss in the next paragraph how this optimization problem can be addressed; then, we provide an introduction to combinatorial optimization methods based on the computation of the minimum cut of an appropriate graph. Sections 4.2.3 and 4.2.4

describe two applications of MRF modeling and these min-cut optimization methods for SAR image regularization.

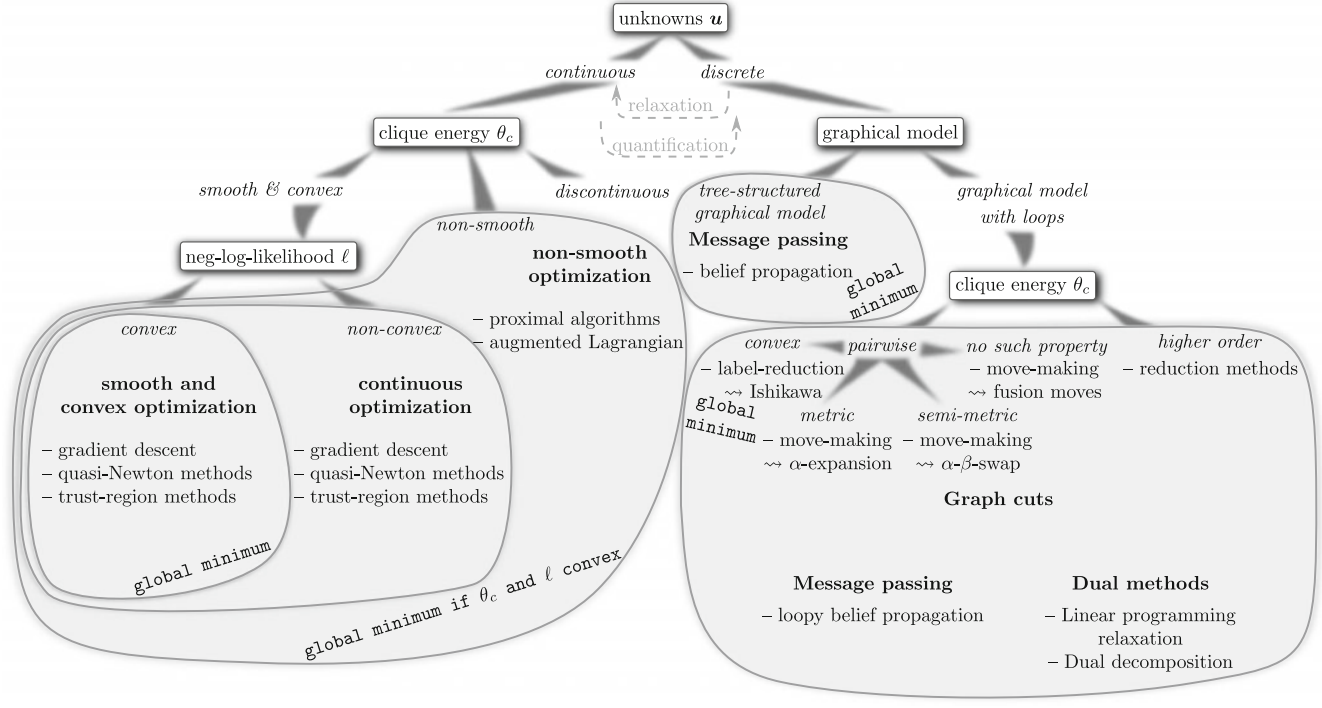
4.2.2.1 Mathematical Optimization Techniques for MAP Estimation

The nature of the minimization problem of Eq. (4.45) involved in the computation of the MAP estimator depends on several factors:

- the domain of the vector of parameters $\mathbf{u}$:
  This domain may be a continuous range (e.g., $u_i \in \mathbb{C}$ or $u_i \in \mathbb{R}^+$) or a discrete set of labels (e.g., $u_i \in \{1, \ldots, L\}$). The former case involves *continuous optimization* while the latter implies a *discrete optimization* (or *combinatorial optimization*). The vector $\mathbf{u}$ may even have some of its elements in a continuous range while the rest are in a discrete set, leading to the so-called *mixed integer programming*.
- the neg-log-likelihood function ℓ:
  Due to the presence of speckle in SAR imaging, the distribution of observations is typically modeled with an heavy-tail. The neg-log-likelihood function ℓ is thus generally *non-convex*.
- the clique energy θ_c:
  Depending on the task at hand (regularization, segmentation, classification), the choice of a particular clique energy θ_c varies. The clique energy may be a smooth function (e.g., $\theta(u_i, u_j) = \beta\,(u_i - u_j)^2$), a non-smooth function (e.g., $\theta(u_i, u_j) = \beta\,|u_i - u_j|$) or a discontinuous function (e.g., $\theta(u_i, u_j) = 0$ if $u_i = u_j$ and $\theta(u_i, u_j) = \beta$ otherwise).

Mathematical optimization is a very wide topic; depending on the domain of the unknown parameters $\mathbf{u}$, the neg-log-likelihood function ℓ and the clique energy θ_c, computation of the MAP estimator involves nonlinear programming methods (for convex or non-convex optimization), non-smooth optimization or combinatorial optimization. It is therefore not surprising that a wide range of algorithms have been developed to address the MAP estimation problem in Markov random fields. While stochastic methods like simulated annealing or Monte Carlo sampling methods have been heavily investigated in the 1980s and 1990s [33], the success of combinatorial optimization methods for MAP inference has driven much of the effort since the end of the 1990s [34]. We report in the scheme on p. 157 some of the main deterministic methods in use to date. Those methods may be broadly separated into continuous optimization methods and combinatorial (i.e., discrete) optimization methods. The nature of the unknowns $\mathbf{u}$ generally dictates the choice of continuous or discrete methods. However, by quantization, unknowns that lie in a continuous range may be turned into discrete unknowns (i.e., taking one of a finite number of states) and combinatorial methods be then applied in order to benefit from theoretical guarantees of identifying the global minimum in some multimodal non-convex cases. Such an approach will be illustrated in Sects. 4.2.3 and 4.2.4. Conversely, a problem originally formulated with discrete

Some optimization methods for MAP inference



unknowns (e.g., $u_i \in \{0, 1\}$) may be turned into a continuous problem by a process called *relaxation* that consists of replacing integral constraints with bound constraints (e.g., $u_i \in [0, 1]$). Efficient continuous optimization methods can then be applied, at the cost of a loss of theoretical guarantees, except for some notable special cases [35]. When both the clique energy and the neg-log-likelihood are smooth (i.e., continuously differentiable), the gradient of the objective function $\mathscr{F}$ provides information to decide for a search direction. Gradient descent methods iteratively improve the current solution $\mathbf{u}^{(k)}$ by a step of length α_k in the direction opposite to the gradient evaluated at $\mathbf{u}^{(k)}$: $\mathbf{u}^{(k+1)} = \mathbf{u}^{(k)} - \alpha_k \cdot \nabla_{\mathbf{u}}[\mathscr{F}_{\mathbf{v}}](\mathbf{u}^{(k)})$. Quasi-Newton methods improve the convergence speed by choosing a more efficient search direction computed with an approximation of the inverse of the Hessian matrix built iteratively from the successive gradient vectors. Trust region methods restrict step sizes within a region in which the second-order approximation of $\mathscr{F}_{\mathbf{v}}$ is considered valid. Provided that step lengths and search directions (or the size of the trust region) are carefully chosen, global convergence[7] to a global minimum can be proven, when $\mathscr{F}$ is convex [36]. When $\mathscr{F}$ is not convex, special care must be taken to ensure that the local quadratic approximations used in quasi-Newton or trust region methods are positive definite. Convergence can then be guaranteed only to a local minimum. It is worth to note that given the large number of unknown parameters generally involved (the dimension of vector $\mathbf{u}$), limited-memory methods like L-BFGS [37] are considered, i.e., the approximation of the Hessian is not explicitly stored but implicitly used based on the last few gradient and iterate values.

Non-smooth objective functions $\mathscr{F}$ are very common owing to the success of sparsity-inducing priors, e.g., clique energies involving L_1 norms. Because of the non-differentiability of $\mathscr{F}$, continuous optimization methods do not directly apply. Several efficient algorithms were derived based on smoothing methods: proximal methods (see, e.g., reviews [38, 39]) or the augmented Lagrangian [40–42]. These methods have guaranteed convergence to a global minimum, provided that $\mathscr{F}$ be convex. Some of these methods also apply to non-convex objective functions $\mathscr{F}$, with the guarantee to converge to a local minimum.

Three main families of methods have been considered to address discrete optimization problems in MAP inference: graph-cuts, message passing algorithms, and dual methods. We devote the next paragraph to the presentation of the principle of graph-cuts and describe in Sects. 4.2.3 and 4.2.4 applications of graph-cuts to MRF models in SAR imaging. Belief propagation (aka min-sum) [43] is a message passing algorithm that leads to a global minimum when the graphical model has a tree structure. When the graph has loops, a variant called *loopy belief propagation* [44] can be applied, without theoretical guarantee of convergence [34]. Finally, several dual methods have been considered. Since the objective function $\mathscr{F}$ writes as a sum of terms, it can be recast as an integer linear program by introducing indicator variables that select the right terms among all possible likelihood terms and clique energies.

[7] By global convergence, it is meant that the algorithm converges even if the initial value $\mathbf{u}^{(0)}$ is far from the optimum.

Since the resulting integer linear program is in general NP-hard to solve, it is relaxed into a linear program, leading to an approximate solution.

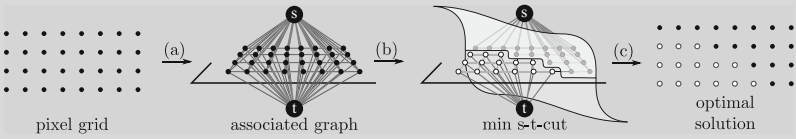
The paper [45] compared 27 different methods on several MAP inference problems with discrete unknowns arising in computer vision. From their analysis, it appears that depending on the structure and size of the problem, graph-cut methods, dual methods, or message passing should be preferred.

4.2.2.2 Minimization Methods Based on Graph-Cuts

This section is intended to introduce the main minimization methods based on the computation of minimum cuts. These methods will form the core of the SAR image processing frameworks described in Sects. 4.2.3 and 4.2.4.

We begin by describing the historically first graph-cut method for MAP inference in a MRF. This method, developed by Greig, Porteous and Seheult [46], applies to binary MRF with pairwise terms that favor neighboring sites to share an identical binary value: (attractive) Ising model.

Finding an optimal binary labeling by graph-cuts: Ising model



To find the optimal binary labeling $\mathbf{u}^\star$, i.e., the solution to the combinatorial problem:

$$\arg\min_{\mathbf{u}\in\{0,1\}^n} \sum_i \mathscr{D}(u_i, v_i) + \sum_{i\sim j} \mathscr{R}(u_i, u_j)\,, \text{ with } \mathscr{R}(u_i, u_j) = \begin{cases} 0 & \text{if } u_i = u_j \\ \beta & \text{if } u_i \neq u_j \end{cases}$$

the method of Greig et al. [46] performs three steps:

(a) it builds a graph with a node for each pixel of the image plus two terminal nodes (the source s and the sink t); edges of the graph connect each node to the two terminal nodes and to their nearest spatial neighbors;
(b) a maximum flow/minimum cut is computed over the graph;
(c) the optimal labeling is derived by analyzing the minimum cut.

Our description only outlined the method. Before justifying that it can actually provide the optimal labeling $\mathbf{u}^\star$, we shall introduce the formal definition of the minimum s-t-cut of a graph.

A *flow network*[8] $\mathcal{G} = (\mathcal{N}, \mathcal{E})$ is a directed graph, defined by a set of nodes $\mathcal{N}$ and a set of directed edges $\mathcal{E}$. Each edge connecting a node x to a node y is given a nonnegative value called *capacity* $\kappa(x, y)$. If two nodes x' and y' are not connected in the graph, we set $\kappa(x', y') = 0$. Two nodes play a special role: the *source* "s" and the *sink* "t". A flow on the graph is a real-valued function $f : \mathcal{N} \times \mathcal{N} \to \mathbb{R}$ that verify the following two constraints:

- *capacity constraint:* $\forall x \in \mathcal{N},\ \forall y \in \mathcal{N},\ 0 \leq f(x, y) \leq \kappa(x, y)$ (the flow never exceeds edge capacities);
- *flow conservation:* $\forall x \in \mathcal{N} \setminus \{\mathrm{s}, \mathrm{t}\},\ \sum_{y \in \mathcal{N}} f(y, x) = \sum_{y \in \mathcal{N}} f(x, y)$ (the flux that enters a node equals the flux that exits that node, there is no accumulation or loss except at the source and sink nodes).

The value $|f|$ of a flow f is defined as the difference between all flux that exits the source and the flux that enters the source:

$$|f| = \sum_{x \in \mathcal{N}} f(s, x) - \sum_{x \in \mathcal{N}} f(x, s) \tag{4.46}$$

The *maximum flow* problem is to find a flow f satisfying the above two constraints and such that its value $|f|$ be maximum.

An *s-t-cut* $(\mathcal{S}, \mathcal{T})$ is a partition of the set of nodes $\mathcal{N}$ into two sets $\mathcal{S}$ and $\mathcal{T}$ such that the source s be in set $\mathcal{S}$ and the sink t in set $\mathcal{T}$. The *cost* (or capacity) $\mathcal{C}$ of the s-t-cut $(\mathcal{S}, \mathcal{T})$ is the sum[9] of the capacities of all edges connecting a node in $\mathcal{S}$ to a node in $\mathcal{T}$:

$$\mathcal{C}(\mathcal{S}, \mathcal{T}) = \sum_{x \in \mathcal{S}} \sum_{y \in \mathcal{T}} \kappa(x, y) \tag{4.47}$$

Max-flow/min-cut theorem [47]: The maximum flow value is equal to the minimum cost[10] among all s-t-cuts ($\max_f |f| = \min_{\mathcal{S}, \mathcal{T}} \mathcal{C}(\mathcal{S}, \mathcal{T})$). Moreover, all edges of the minimum cut are saturated, the minimum cut can thus be readily obtained from a maximum flow f by exploring the *residual graph*, i.e., the graph obtained by keeping all nodes in $\mathcal{N}$ and, for each edge $x \to y$, by creating a pair of edges $x \to y$ and $y \to x$ with a *residual capacity* $\kappa_r(x, y) = \kappa(x, y) - f(x, y)$ in the $x \to y$ direction and a capacity $f(x, y)$ in the $y \to x$ direction [48]. The set $\mathcal{S}$ is formed by all nodes that can be reached in the residual graph from the source s.

There are several approaches to compute the maximum flow of a flow network: (i) *augmenting path* algorithms that maintain a valid flow and search for paths in the residual graph; (ii) *push-relabel* methods that maintain a *preflow*, i.e., a flow that respects the capacity constraints but may accumulate some flux at some nodes

[8] Also called an s-t-graph.

[9] Note that edges connecting a node in $\mathcal{T}$ to a node in $\mathcal{S}$ do not enter the sum.

[10] The maximum flow and the minimum cut problems can be formulated as two dual linear programs.

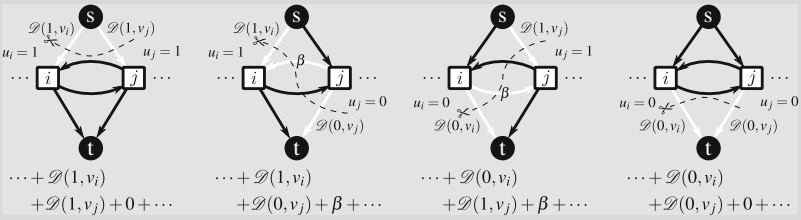
(no flow conservation); (iii) Hochbaum's pseudo-flow variants, i.e., methods that consider a flow that respects the capacity constraints but not the flow conservation (accumulation or deficit at some nodes). Several variants of each category have been compared experimentally on computer vision problems in [49, 50]. Kolmogorov's implementation [51] of Boykov and Kolmogorov algorithm [49] is probably the most widespread to date.

Now that we have formally defined the minimum cut of graphs, we can return to the graph-cut method for MAP estimation an Ising model and describe how to set the capacities of edges in the associated graph described on p. 23.

Finding an optimal binary labeling by graph-cuts: Ising model (cont.)

A cut separating the graph described on p. 23 into a partition containing the source s and another containing the sink t necessarily cuts, for each node, either the edge connecting the node to the source or the edge connecting the node to the sink. The set of all possible cuts is therefore in bijection with the set of all possible binary labelings $\mathbf{u} \in \{0, 1\}^n$.

Capacities of the edges are set so that each of the 4 possible cuts that would disconnect a portion of the graph containing two neighboring nodes i and j has costs matching the cost function in Ising model:



In words, the node associated with the ith pixel is connected by a downward arc to the source, with capacity $\mathscr{D}(1, v_i)$, and to the sink by an arc with capacity $\mathscr{D}(0, v_i)$. Furthermore, neighbor pixels i and j are represented in the graph by a pair of arcs connecting nodes i and j, each with a capacity β.

Beyond the important but restrictive Ising model, one may wonder if similar graph constructions enjoy the property that all possible cut costs are in bijection with all possible values of the objective function. In their seminal paper [52], Kolmogorov and Zabih show that a class of pairwise[11] binary energy minimization problems can be solved by graph-cut:

$$\underset{\mathbf{u}\in\{0,1\}^n}{\arg\min} \quad \sum_i \mathscr{D}(u_i, v_i) + \sum_{i\sim j} \mathscr{R}(u_i, u_j) \tag{4.48}$$

[11] The class of MRF with cliques involving triplets is also covered in [52].

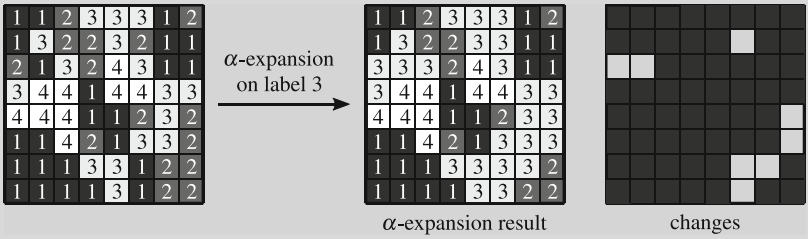
with $\mathscr{R}$ a *submodular* function, i.e., a function such that

$$\mathscr{R}(0,0) + \mathscr{R}(1,1) \leq \mathscr{R}(0,1) + \mathscr{R}(1,0) \tag{4.49}$$

Solving the minimization problem for binary functions $\mathscr{R}$ that are *not* submodular is in general NP-hard [52]. Solutions can still be obtained by graph-cuts, but they are only approximate [53].

The extension of graph-cut methods to multi-label MRFs (i.e., $\mathbf{u} \in \{1, \ldots, L\}^n$) can be done along two different paths: *move-making* methods and *label-reduction*. Move-making are greedy approaches based on a sequence of binary subproblems. At each step, the current solution is improved by finding the best solution within a subset of the whole domain $\{1, \ldots, L\}^n$ that includes the current solution. We describe below the three most important moves.

Move-making graph-cut algorithms: α-expansion [54]



In the α-expansion algorithm, the elementary move considers which combination of pixels should optimally be either kept to their previous value or changed to the new proposal α. Those pixels are marked as "changed" in the binary image on the right-hand side of the illustration. Identifying those pixels is a binary labeling problem which can thus be exactly solved, provided that the submodularity constraint is fulfilled:

$$\forall \alpha,\ \forall b,\ \forall c,\ \mathscr{R}(b,c) + \mathscr{R}(\alpha,\alpha) \leq \mathscr{R}(b,\alpha) + \mathscr{R}(\alpha,c)$$

which is guaranteed when the regularization is a *metric*, i.e., such that:

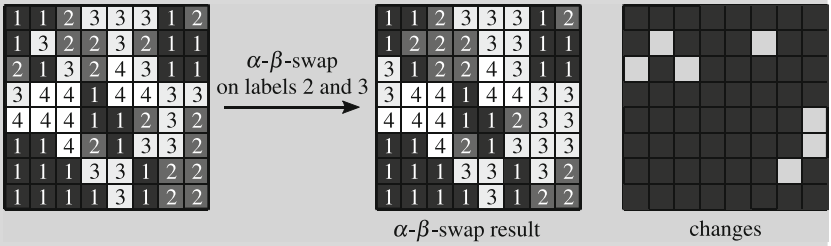
$$\forall b,\ \forall c, \quad \mathscr{R}(b,c) = 0 \ \Leftrightarrow\ b = c \tag{4.50}$$

$$\forall b,\ \forall c, \quad \mathscr{R}(b,c) = \mathscr{R}(c,b) \geq 0 \tag{4.51}$$

$$\forall \alpha,\ \forall b,\ \forall c, \quad \mathscr{R}(b,c) \leq \mathscr{R}(b,\alpha) + \mathscr{R}(\alpha,c) \tag{4.52}$$

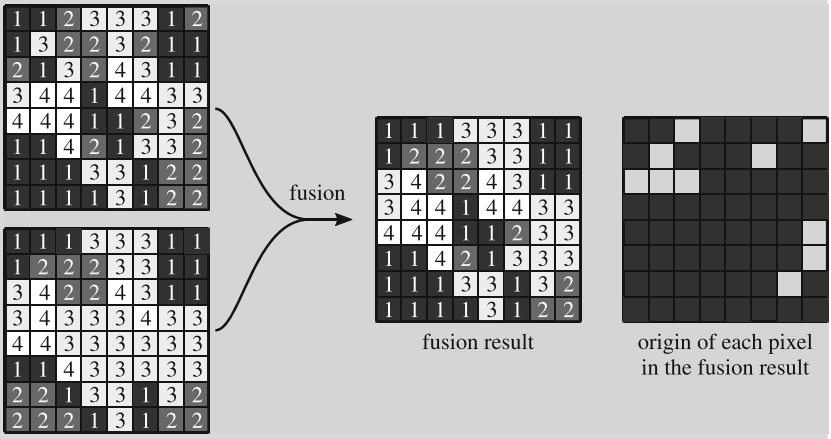
The algorithm performs a sweeping of all labels 1, 2, …, L, each time performing an α-expansion on the current label, until stability.

Move-making graph-cut algorithms: α-β-swap [54]



In the α-β-swap algorithm, only pixels with labels α or β may change by possibly swapping their labels. Finding the optimal labeling problem again boils down to a binary labeling task that can be solved exactly, provided that $\mathscr{R}$ is a *semi-metric*, i.e., verifies Eqs. (4.50) and (4.51). The algorithm performs a sweeping of all pairs of labels until stability.

Move-making graph-cut algorithms: fusion move [55]



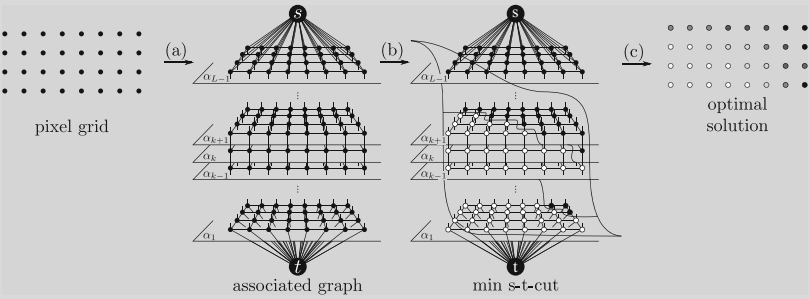
The fusion move generates an improved solution $\mathbf{u}^{(k+1)}$ by fusion of two alternative approximative solutions $\mathbf{u}_1^{(k)}$ and $\mathbf{u}_2^{(k)}$. Deciding which combination of sites should optimally be given the labels from $\mathbf{u}_1^{(k)}$ and which should keep labels from $\mathbf{u}_2^{(k)}$ is a binary labeling problem. Since the submodularity condi-

tion is in general not fulfilled, an approximate solution of that binary problem is computed using an algorithm such as in [53].

Several ways to compute the two proposals $\mathbf{u}_1^{(k)}$ and $\mathbf{u}_2^{(k)}$ can be considered. An approach called *log-cuts* [56] has been proposed to accelerate the α-expansion algorithm by sequentially finding each bit of a binary coding of the labels as $\lceil \log L \rceil$-bit long words instead of considering all L labels at each sweep. Another use of fusion moves is to merge solutions obtained by a continuous optimization method [55].

The most prominent algorithm of the class of *label-reductions* is due to Ishikawa [57]. It provides the global optimum even if $\mathscr{D}$ is not convex, at the cost of building a large graph with $n \times L$ nodes. We describe here the constructed graph when the regularization $\mathscr{R}$ takes the special form[12]: $\mathscr{R}(u_i, u_j) = |u_i - u_j|$.

Label-reduction with Ishikawa graph for exact minimization of convex pairwise MRFs



The scheme illustrates the constructed graph to minimize an energy of the form:

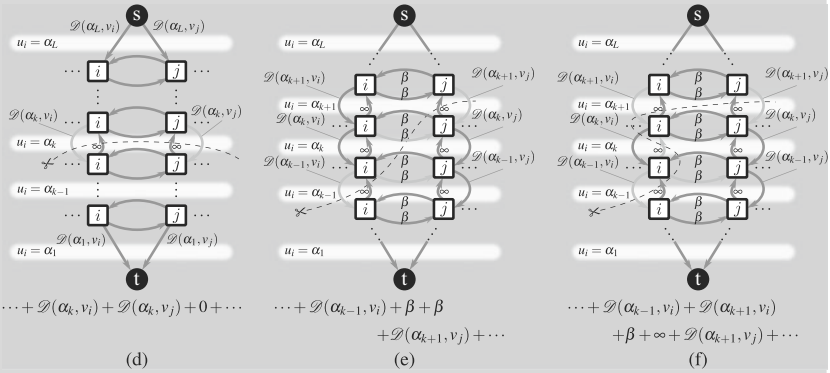
$$\underset{\mathbf{u}\in\{\alpha_1,\ldots,\alpha_L\}^n}{\arg\min} \quad \sum_i \mathscr{D}(u_i, v_i) + \beta \sum_{i\sim j} |u_i - u_j| \tag{4.53}$$

where $\mathscr{D}$ may be non-convex and the second sum corresponds to the anisotropic total variation.

Like for Ising model, the method proceeds in three steps: (a) a graph is associated with the pixel grid on which the MRF is defined; the graph has L layers of n nodes and each node is connected to its spatial neighbors within the layer and to its two corresponding nodes in the preceding and following layers; (b) a minimum s-t-cut is computed; (c) the height (i.e., layer) at which each column of node is cut defines the optimal label in the resulting solution.

[12] The method described in [57] applies to all convex pairwise terms $\mathscr{R}$, see also [58, 59].

We now illustrate how capacities are set to each arc so that cut costs are in bijection with energies of the fields **u**.



Downstream arcs of the ith column of nodes are given a capacity corresponding to values $\mathscr{D}(u_i, \alpha_L)$ to $\mathscr{D}(u_i, \alpha_1)$ (subfigure (d)). Arcs that connect spatial neighbors, i.e., arcs with a given layer, all have a capacity β. As illustrated in subfigure (e), when two neighboring sites take different labels the cut necessarily includes as many arcs with weight β as label difference between sites, thereby representing the energy of Eq. (4.53). Finally, upstream arcs with infinite capacity are introduced to prevent from cutting twice a given column of nodes (subfigure (f)). With this construction, all cuts with *finite cost* are in bijection with all labelings of the MRF.

4.2.3 *Application to SAR Image Denoising: Joint Regularization and Fusion with Optical Data*

In this section, we focus on the application of the MRF formalism to the regularization of SAR images and illustrate how the graph-cut methods described previously can apply to speckle reduction.

Amplitude regularization. Due to speckle phenomenon, the amplitude of a SAR image suffers from strong fluctuations from a pixel to another. Hence, the magnitude of the spatial gradient is large in most of the regions of the image (except dark areas due to shadows or smooth surfaces that diffuse most of the incoming SAR pulse in the reflection direction, away from the radar antenna) as shown in Fig. 4.5(a). As shown in Sect. 4.1.3, multi-looking done by incoherent averaging of SAR intensity within a window reduces speckle fluctuations at the cost of a loss of resolution. Figure 4.5(b) shows the magnitude of the spatial gradient applied on a 100-looks image, i.e., a

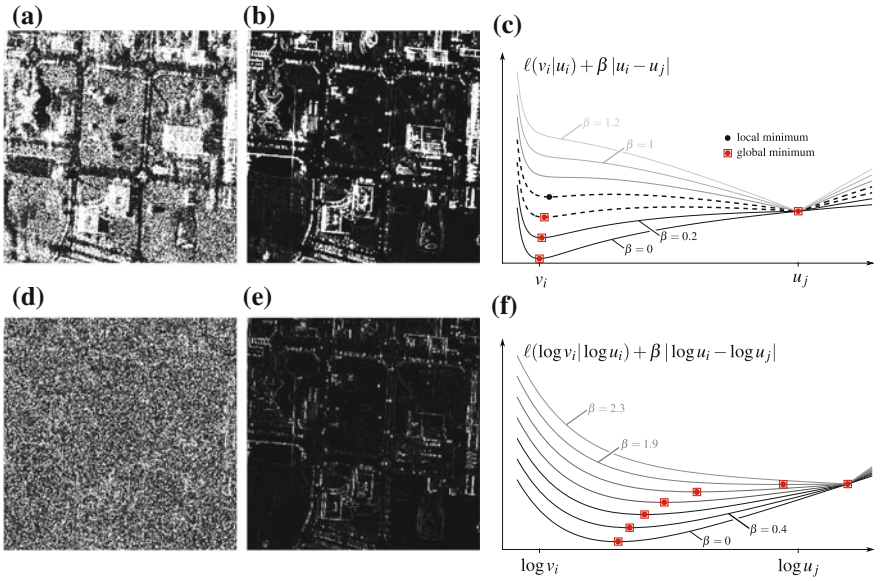


Fig. 4.5 SAR amplitude regularization by total-variation (TV) minimization: **a** the magnitude of the spatial gradient of a single-look amplitude image is large and nonzero almost everywhere while **b** the magnitude of the spatial gradient is much sparser on a 100-looks image; **c** the minimization problem for TV minimization of the amplitude is non-convex. Similarly, **d** the magnitude of the spatial gradient of the logarithm of the amplitude is nowhere zero with a single look, while **e** a 100-looks log-amplitude image has a very sparse gradient; **f** TV of the log-amplitude is convex, in contrast to **c**

very-high-resolution image whose resolution was reduced by 10 in each direction in order to mitigate almost all the speckle effect. It can be noticed that this radar scene (suburban area of the city of Toulouse, France, at 1-m spatial resolution © CNES & ONERA) has a majority of spatial gradients close to zero. It is thus natural with Markov random field framework to restore the amplitude (i.e., reduce speckle) by introducing a sparse gradient prior: the MAP estimator amounts to find a trade-off between data fidelity (accounted for via a data-fitting term derived from speckle models) and regularity (measured via the total-variation semi-norm):

$$\arg\min_{\mathbf{u}} \quad \sum_i \ell(v_i|u_i) + \beta \sum_{i\sim j} w_{i,j}|u_i - u_j| \tag{4.54}$$

where $\mathbf{u}$ is the image of restored amplitude (i.e., square root of the reflectivity $u_i = \sqrt{\mathrm{R}_i}$), $\mathbf{v}$ is the speckle-corrupted amplitude image ($v_i = \mathrm{A}_i$), the notation $i \sim j$ indicates that i and j are two pixels that are neighbors (i.e., they define clique of order two on the graph associated with the MRF), β is a regularization weight, and $w_{i,j}$ is used to weight differently pixels that are separated by a different distance (i.e., left-right-top-bottom neighbors vs diagonal neighbors). The neg-log-likelihood

function for SAR amplitude in L-looks images is $\ell(v_i|u_i) = L\left(\frac{v_i^2}{u_i^2} + 2\log u_i\right)$ under the assumption of fully developed and uncorrelated speckle (see the box in Sect. 4.1.3.4). Since this data term ℓ is not convex, the MAP estimation problem is more challenging than usual total-variation minimization problems [60–62]. We illustrate in Fig. 4.5(c) this non-convexity. When the regularization is either very weak or very strong, the posterior energy, though non-convex, becomes unimodal (in this simple example involving a single pixel i and its neighbor j with given value u_j). However, for intermediate regularization values (shown with a dashed curve), there are local minima. After quantization, this non-convex problem can be solved using the graph-cuts construction shown p. 28, or an approximate minimization based on move-making techniques [61].

As presented in Sect. 4.1.3, the logarithm of the amplitude of a speckle-corrupted SAR image displays a stationary variance (see Fig. 4.5(d)), a phenomenon called variance stabilization. While Gamma-distributed fluctuations are signal-dependent (their variances depend on the underlying reflectivity), this transform turns the fluctuations into an additive signal-independent perturbation with fixed variance $\sigma^2 = \Psi(1, L)$. Again, the magnitude of the spatial gradient of a log-transformed speckle-free 100-looks image is very sparse: mostly zero (or close to zero), which justifies following a total-variation (TV) minimization approach on the log-transformed values $\tilde{v}_i = \log v_i$ (and $\tilde{u}_i = \log u_i$). Considering the TV minimization in the log-domain has two advantages: (i) Speckle is reduced equally in all regions (since the variations in the log-transformed image are spatially homogeneous as shown in Fig. 4.5(d)) and (ii) the problem becomes convex, as illustrated in Fig. 4.5(f). Nevertheless, this procedure is known to be biased if a quadratic likelihood term is used as the log transform fluctuations do not exactly follow a Gaussian distribution [63]. In this case, a bias correction should be applied, leading to the inversion formula: $u_i = \exp(\tilde{u}_i + \log L - \Psi(L))$. If the adapted likelihood is used (based on the Fisher–Tippett distribution of $\tilde{u}_i$), no bias correction is needed. The homomorphic approach has been followed by several authors [64–66].

Beyond regularization: joint estimation of the background and detection of strong scatterers. A known drawback of TV minimization is the tendency to remove isolated points, or to merge close pixels that have comparable values. To counterbalance this drawback, Çetin et al. [67, 68] used both a total-variation term and a sparsity-promoting term to favor isolated scatterers. Such an approach can be pushed further by decomposing the radar scene into two components: a background (with low total variation) and strong scatterers (with low ℓ_0 pseudo-norm) [27, 69].

Interferometric phase regularization. After proper regularization/correction for plane Earth fringes, interferometric phase provides information on the elevation of the structures of the radar scene. However, because of the strong fluctuations due to speckle, this elevation information is too noisy to be directly usable. After multi-looking (i.e., averaging in a small local neighborhood), phase can be further regularized following a MRF approach. The likelihood after multi-looking is well approximated by a (non-stationary) Gaussian model with a variance that depends

on the local coherence [61]. Since in urban areas, the interferometric phase displays strong discontinuities, total variation is once again a natural prior.

Joint regularization: contour co-localization. Rather than performing the regularization of the interferometric phase independently of the amplitude, it is beneficial to jointly regularize the two channels and favor the co-localization of edges. This can be performed by penalizing the maximum magnitude of the two spatial gradients (with proper scaling of the two images) [61]:

$$\theta_{i,j}(u_i^a, u_i^\varphi, u_j^a, u_j^\varphi) = w_{i,j} \max(|u_i^a - u_j^a|, |u_i^\varphi - u_j^\varphi|) \tag{4.55}$$

with u_i^a the regularized amplitude at pixel i and u_i^φ the regularized interferometric phase at the same pixel.

Fusion of edge information from an optical image. Optical images of the region of interest are often available. It may be desirable to achieve a form of fusion of optical and SAR information. More details on data fusion methods will be given in Chaps. 6 and 7. A critical issue for optical/SAR fusion is that of registration of the optical and SAR images since projecting from SAR geometry to optical geometry requires knowledge of all sensor parameters and an estimate of the height at each pixel of the SAR image. With SAR interferometry, this can be achieved. In [70], it was proposed to extract edges from the optical image,[13] and feed this information into a conditional random field, namely by favoring edges in the regularized SAR image where edges had been found in the optical image:

$$\theta_{i,j}(u_i^a, u_i^\varphi, u_j^a, u_j^\varphi) = w_{i,j} \max(0, 1 - \alpha_{\text{opt}}|u_i^o - u_j^o|) \max(|u_i^a - u_j^a|, |u_i^\varphi - u_j^\varphi|) \tag{4.56}$$

where u_i^o is the (independently regularized) optical image intensity at pixel i and α_{opt} is a parameter to control the weight of the optical image (setting α_{opt} to zero cancels all effect of the optical image). Figure 4.6 illustrates the principle of this optical/radar fusion approach.

4.2.4 Application to Phase Unwrapping of Multi-channel Interferometry

SAR interferometric (InSAR) systems allow to reconstruct height profile starting from the measure of the interferometric phase, using the relation (4.25). Considering that the interferometric phase is measured in the principal interval $[-\pi.\pi)$, the relation between height z and phase ϕ, in the absence of noise, can be written as [2]:

[13] A total-variation regularization was applied on the optical image to remove low-significance textures.

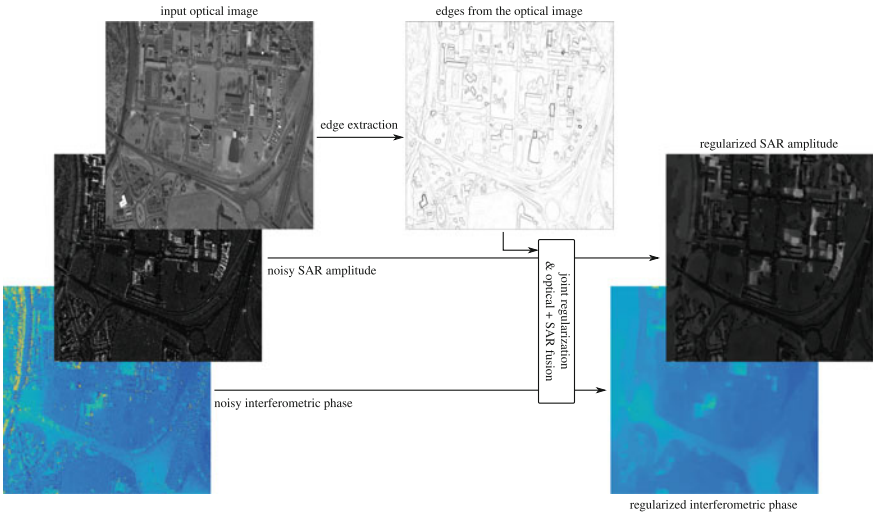


Fig. 4.6 Illustration of joint amplitude/interferometric phase regularization with information from an additional optical image

$$\phi = \langle \frac{4\pi B \cos(\theta - \sigma)}{\lambda r \sin\theta} z \rangle_{2\pi} \tag{4.57}$$

where $\langle \alpha \rangle_{2\pi}$ denotes α modulo 2π evaluation. The relation is strictly nonlinear and for retrieving the height, the absolute value of the phase has to be restored. This problem is commonly known has *phase unwrapping* (PhU) [9]. This is, however, an ill-posed problem, since it admits an infinite number of solutions, if no further information is added. A regularization based on physical considerations is needed to solve the problem.

In literature in the last decades, several approaches have been proposed to solve PhU problem. These approaches mainly belong to three classes: *path following*, *minimum norm*, and *statistical estimation* approaches.

Most PhU algorithms are effective only in case the *Itoh condition* [71] is satisfied. This condition implies that the maximum phase difference between neighboring pixels is smaller than π. In this case, the absolute phase can be easily determined, up to a constant. Unfortunately, the condition is often not met in InSAR systems mainly due to the presence of large height discontinuities in the observed scene or due to the presence of the interferometric noise. In both cases, PhU operation becomes a difficult task.

In literature, the expression *multi-channel phase unwrapping* (MCPhU) is adopted to refer to a set of techniques able to solve PhU problem by exploiting several interferometric data of the observed scene [72–75]. MCPhU is able to overcome limitation of single-interferogram approaches and to retrieve the height of the scene even in case Itoh condition is not met.

Multi-channel interferograms can be obtained using mainly two different approaches: multi-baseline and multi-frequency acquisitions. In the former, the scene is observed by more than two SAR systems (at least three) from slightly different positions (i.e., with different baselines). We recall that the distance between SAR sensors has been previously defined as baseline (Sect. 4.1.2). The latter, multi-frequency, consists in using SAR sensors operating at different frequencies and/or by partitioning the Fourier spectrum in non-overlapped sub-bands.

Starting from the relation (4.57), in case of K-independent interferometric channels, the interferometric measured phase signal for the channel k can be written as:

$$\phi_k = \langle \upsilon_k z \rangle_{2\pi} \tag{4.58}$$

where υ_k is defined as:

$$\upsilon_k = \frac{4\pi B_k \cos(\theta - \sigma)}{\lambda r \sin\theta} \tag{4.59}$$

$$\upsilon_k = \frac{4\pi B \cos(\theta - \sigma)}{\lambda_k r \sin\theta} \tag{4.60}$$

in case of multi-baseline and in case of multi-frequency, respectively. For the rest of the chapter, υ_k will be considered without specifying the type of diversity (i.e., baseline or frequency), since the approach is mainly the same.

An effective way to combine the different available interferograms is to use statistical estimation theory, in particular exploiting MAP approach (Sect. 4.2.2). Let us start from MAP solution of Eq. (4.45) and let us characterize the parameters for the specific case of MCPhU. Let us start from the likelihood term. The likelihood function for SAR interferometry in case of single channel for pixel i (Sect. 4.1.3) is:

$$p(\phi_{k,i}|\rho_{k,i}, z_i) = \frac{1}{2\pi} \frac{1-\rho_{k,i}^2}{1-\rho_{k,i}^2 \cos(\phi_{k,i} - \upsilon_k z_i)^2} \left(1 + \frac{\rho_{k,i} \cos(\phi_{k,i} - \upsilon_k z_i) \cos^{-1}(-\rho_{k,i} \cos(\phi_{k,i} - \upsilon_k z_i))}{(1-\rho_{k,i}^2 \cos(\phi_{k,i} - \upsilon_k z_i)^2)^{1/2}}\right) \tag{4.61}$$

In case of multi-channel systems, exploiting statistical independence between interferograms [76] and supposing the coherence values $\rho_{k,i}$ are given,[14] the negative log-likelihood function can be written as:

$$\ell((\phi_{k,i})_{\{k\}}|z_i) = -\sum_k \log p(\phi_{k,i}|\rho_{k,i}, z_i) \tag{4.62}$$

[14] In practice, they are empirically estimated using local samples and Eq. (4.35).

Moving to the a priori, different models can be used. In particular, one of the most effective is the so-called local Gaussian MRF, which is a Gaussian MRF with locally defined hyperparameters [77]. These hyperparameters, once estimated, can help in properly tuning the MRF, by adapting it to the local behavior of the scene under investigation [78].
Another effective model is to adopt the previously defined TV model (Sect. 4.2.3). Using TV, the a priori can be written as $\sum_{i\sim j} w_{i,j}|z_i - z_j|$. Thus, the MAP estimator for the height estimation is given by:

$$\arg\min_{\mathbf{z}} \sum_i \ell((\phi_{k,i})_{\{k\}}|z_i) + \beta \sum_{i\sim j} w_{i,j}|z_i - z_j| \tag{4.63}$$

The MCPhU with total-variation (MCPhU-TV) approach [79] uses Ishikawa method (Sect. 4.2.3) for the optimization step of Eq. (4.63). Results of MCPhU-TV applied to two different datasets (natural and urban scenario) are reported in Figs. 4.7 and 4.8.

An important point has to be highlighted when dealing with MCPhU approaches. In order to correctly combine the different available interferograms and to restore the correct relation between them, the so-called phase offset compensation procedure

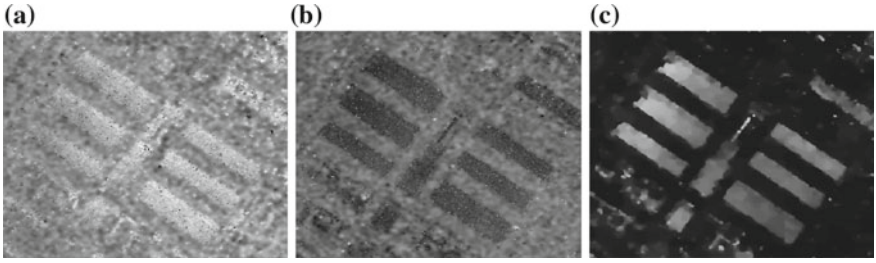


Fig. 4.7 MCPhU-TV applied to urban scenario using six TerraSAR-X X-band multi-baseline interferograms (under ISPRS-DEM Project) of Barcelona (Spain): **a** and **b** two of the available interferograms; **c** 3D reconstruction of the observed scene using MCPhU-TV

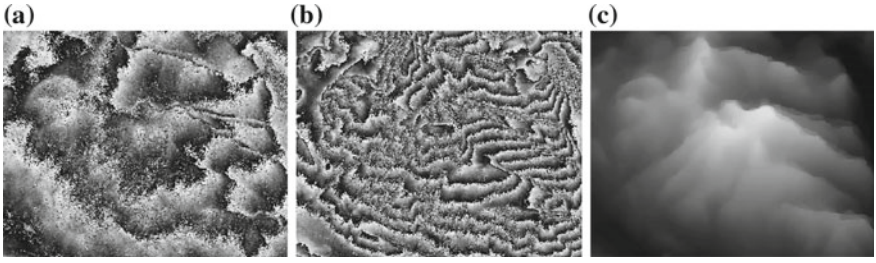


Fig. 4.8 MCPhU-TV applied to natural scenario using six ERS C-band multi-baseline interferograms of Serre-Ponçon (France): **a** and **b** two of the available interferograms; **c** 3D reconstruction of the observed scene using MCPhU-TV

has to be carried out. These offsets arise as a combination of several factors such as processing errors, atmospheric effects, and parallel-baseline uncertainty [80]. Different techniques have been proposed in literature to estimate and compensate such offsets [81, 82].

An evolution of MCPhU-TV in the Markovian-MAP framework, based on the joint exploitation of both amplitude and phase of the available complex data, able to unwrap and simultaneously regularize the observed data has been proposed in [83], named multi-channel phase and amplitude regularization (MCPAR). In particular, MCPAR, by means of the exploitation of amplitude data within the unwrapping chain as in Eq. (4.55), allows a better preservation of height discontinuities.

4.3 Patch-Based Models for SAR Imagery

Section 4.2 described the benefit of capturing the statistical dependency of the spatial distribution of physical parameters in a SAR image. In this section, we describe two alternatives based on patches. Formally, a patch $\mathbf{u}_{\square p}$ is a m-dimensional vector obtained by restricting the image $\mathbf{u}$ to the pixels located within a rectangular window centered at pixel p. The square in the notation $\mathbf{u}_{\square p}$ is used to emphasize that the dimension m of the patch is smaller than the dimension n of the whole image $\mathbf{u}$, typically m ranges from 3×3 to 15×15 pixels. While MRFs generally model conditional dependencies within a local neighborhood, patches capture the configuration of more extended neighborhoods, thereby encoding potentially complex geometrical and textural features. The first strategy uses a prior model for the distribution of patches and relies on the assumption that patches may have a compact representation. The second one is free of prior modeling and simply reenforces the likelihood by artificially increasing the number of local observations based on similar patches.

4.3.1 From Local Neighborhoods to Patches

An image or a collection of images (of natural scenes) always presents several repeating patterns such as edges, corners, crosses. While it is challenging to model directly the distribution of large images, patches present much less variability because of their smaller size and essentially because of this redundancy principle. As a consequence, the distribution of these patches can be well represented with a statistical model of low complexity. In particular, this opens the way toward statistical learning in order to fit such a distribution on a large dataset of patches.

By defining the prior on large cliques of the size of a patch (i.e., typically involving a hundred of pixels), variational patch-based approaches model the prior distribution of $\mathbf{u}$ in terms of the prior distribution of all its patches:

$$\mathrm{p}(\mathbf{u}) = \frac{1}{\mathscr{Z}} \exp\Big\{ - \sum_{p \in \mathscr{P}} \theta_p(\mathbf{u}_{\square p}) \Big\} \tag{4.64}$$

where $\mathscr{Z}$ is a partition function (see Eq. (4.41)), $\mathscr{P}$ is the set of all central pixels of the image patches, and the functions $\theta_p(\cdot)$ are the *patch energies*.

Similarly to the inference in MRFs, injecting this prior into the MAP estimation requires to solve the following optimization problem:

$$\widehat{\mathbf{u}}^{(\mathrm{MAP})} \in \arg\min_{\mathbf{u}} \mathscr{F}_{\mathbf{v}}(\mathbf{u}) \,, \quad \text{with } \mathscr{F}_{\mathbf{v}}(\mathbf{u}) = \sum_i \ell(v_i | u_i) + \sum_{p \in \mathscr{P}} \theta_p(\mathbf{u}_{\square p}) \tag{4.65}$$

4.3.1.1 Choice of Patch Energies

A simple approach introduced in [84] consists in modeling patches in natural image as a *sparse* combination of elements (the so-called *atoms*) from a dictionary of K patches $\mathbf{D} = (\mathbf{d}_1, \ldots, \mathbf{d}_K)$, where the atoms $\mathbf{d}_k$ of the dictionary are all m-dimensional normalized vectors ($\forall k$, $\|\mathbf{d}_k\| = 1$). The *sparse synthesis* prior considers that a dictionary $\mathbf{D}$ can be found such that all probable patches in natural images can be synthesized as a linear combination of a few atoms and that the fewer atoms used in the combination, the more probable is the patch:

$$\theta_p(\mathbf{u}_{\square p}) = \min_{\mathbf{a}} \|\mathbf{a}\|_0 \quad \text{subject to} \quad \mathbf{u}_{\square p} = \sum_k \mathrm{a}_k \mathbf{d}_k \tag{4.66}$$

where $\|\cdot\|_0$ is referred to as the ℓ_0 pseudo-norm counting the number of nonzero elements of its input vector. Inspired from a large literature regarding wavelet shrinkage techniques, the authors of [84] have suggested using a dictionary whose atoms form a redundant family, i.e., with K larger than m (typically $K = 256$ for patches of size $m = 8 \times 8$). The atoms are then necessarily linearly dependent which has been shown to favor sparsity and reconstruction accuracy. More importantly, the authors have suggested to adapt the dictionary to the noisy image $\mathbf{v}$ itself, leading to the K singular value decomposition (KSVD) formulation:

$$\widehat{\mathbf{u}}^{(\mathrm{KSVD})} \in \arg\min_{\mathbf{u}} \sum_i \ell(v_i | u_i) + \min_{\mathbf{D}} \sum_{p \in \mathscr{P}} \left(\min_{\mathbf{a}} \beta \Big\| \mathbf{u}_{\square p} - \sum_k \mathrm{a}_k \mathbf{d}_k \Big\|_2^2 + \|\mathbf{a}\|_0 \right) \tag{4.67}$$

where $\beta > 0$ is a relaxation parameter of the constraint in (4.66). Minimizing (4.67) is very challenging since the joint minimization on $\mathbf{D}$ and $\mathbf{a}$ is a combinatorial non-convex problem which requires the use of suboptimal optimization strategies. In practice, the minimization is performed alternatively in two steps. In the *sparse coding* step, $\mathbf{a}$ is computed by orthogonal matching pursuit techniques [85]. In the *dictionary update* step, the dictionary $\mathbf{D}$ is updated by using K-truncated singular

value decomposition (SVD). The minimization with respect to $\mathbf{u}$, for a given dictionary and decomposition of all patches, is usually much simpler; it is often performed only once, after $\mathbf{a}$ and $\mathbf{D}$ have been determined, but repeating the whole procedure several times improves the solution [86].

Rather than *synthesizing* the patches, an alternative approach, named *Fields of Experts* (FoE) [87, 88] or *sparse analysis prior* [89], models the distribution of filtered versions of the patches, using different filters (called *experts*):

$$\theta_p(\mathbf{u}_{\square p}) = \sum_{k=1}^{K} \psi\left(\mathbf{f}_k^{\mathrm{t}}\, \mathbf{u}_{\square p};\ \mathrm{a}_k\right) \tag{4.68}$$

where K is the number of experts and the function ψ is chosen to favor sparsity, for example, the ℓ_1 norm (or a smoothed version of the ℓ_1 norm), or a non-convex function like the neg-log-likelihood of Student distribution [88]:

$$\psi\left(\mathbf{f}_k^{\mathrm{t}}\, \mathbf{u}_{\square p};\ \mathrm{a}_k\right) = \mathrm{a}_k \cdot \log\left(1 + \frac{1}{2}\left(\mathbf{f}_k^{\mathrm{t}}\, \mathbf{u}_{\square p}\right)^2\right) \tag{4.69}$$

The filters $\mathbf{f}_k$ are m-dimensional vectors *learned* from a large basis of natural images and are generally high frequency (derivative-like). Weights a_k balance the importance of each expert and are learned jointly with the filters $\mathbf{f}_k$. Given the difficulty of the learning step, only small patches are considered: 5×5 patches in [88], 7×7 patches in [90].

A more recent alternative, inspired from [91], is to model the prior distribution of patches with a Gaussian mixture model (GMM) as:

$$\theta_p(\mathbf{u}_{\square p}) = -\log \sum_{k=1}^{K} \mathrm{w}_k \cdot \mathcal{N}(\mathbf{u}_{\square p};\ \boldsymbol{\mu}_k,\ \boldsymbol{\Sigma}_k) \tag{4.70}$$

where $\mathcal{N}(\mathbf{u}_{\square p};\ \boldsymbol{\mu}_k,\ \boldsymbol{\Sigma}_k) = \dfrac{1}{\sqrt{2\pi}\,|\boldsymbol{\Sigma}_k|^{1/2}} \exp\left(-\frac{1}{2}\left(\mathbf{u}_{\square p} - \boldsymbol{\mu}_k\right)^{\mathrm{t}}\, \boldsymbol{\Sigma}_k^{-1}\left(\mathbf{u}_{\square p} - \boldsymbol{\mu}_k\right)\right)$

leading to the expected patch log-likelihood (EPLL) formulation [92]:

$$\widehat{\mathbf{u}}^{(\mathrm{EPLL})} \in \arg\min_{\mathbf{u}} \left\{ \sum_i \ell(v_i|u_i) + \sum_{p \in \mathscr{P}} \min_{\mathbf{z}_\square} \left(\beta \left\|\mathbf{u}_{\square p} - \mathbf{z}_\square\right\|_2^2 + \theta_p(\mathbf{z}_\square)\right) \right\} \tag{4.71}$$

with $\beta > 0$. As for KSVD algorithm, the relaxation term $\beta \left\|\mathbf{u}_{\square p} - \mathbf{z}_\square\right\|_2^2$ is introduced in order to apply an alternating minimization strategy with respect to unknowns $\mathbf{z}_\square$ and $\mathbf{u}$. While the minimization with respect to the unknown image $\mathbf{u}$ is typically smooth and rather simple, the optimization for $\mathbf{z}_\square$ is much more intricate as the energy $\theta_p(\cdot)$ is non-convex. This latter optimization problem is usually performed

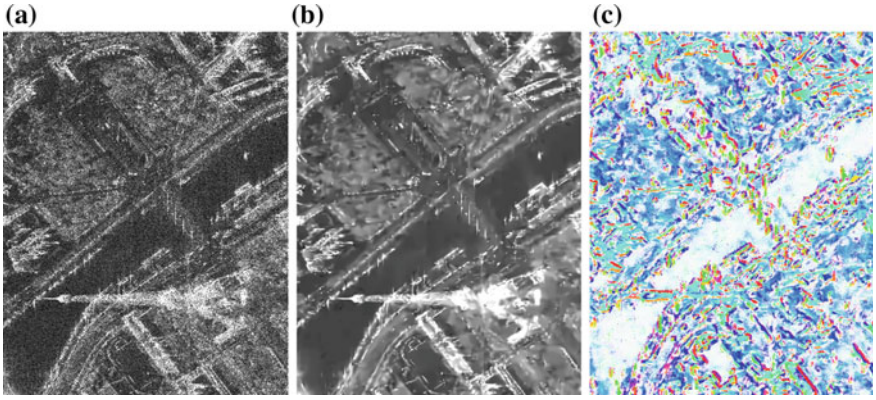


Fig. 4.9 **a** Original single-look intensity TerraSAR-X image (© DLR under project LAN-1746) of Paris (France). **b** Estimation of the reflectivity by SAR EPLL [94]. **c** Illustration of the local adaptivity of EPLL that uses different Gaussian models at each pixel location. Each color represents a Gaussian model, such that similar colors correspond to similar Gaussian models and the lower the saturation the more equally likely the models are

by approaching Eq. (4.70) with only one of the Gaussian models. In the seminal work of [92], this selection is performed for each patch and is updated at each iteration. Unlike the KSVD approach that adapts the dictionary $\mathbf{D}$ to the content of each image, EPLL learns the GMM offline from an external dataset of patches using the classical expectation–maximization technique [93] (some details on this technique are recalled in Chap. 9).

Figure 4.9 provides illustrations of the performance and local adaptivity of an extension of EPLL for SAR intensity images [94] where the likelihood term is chosen to fit the statistics of speckle as described in the Sect. 4.1.3 and below.

4.3.1.2 Choice of Likelihood Term

Variational patch-based approaches are usually defined with the square difference fidelity term $\ell(v_i|u_i) = \frac{1}{2\sigma^2}(v_i - u_i)^2$, known to capture the statistics of additive white Gaussian noise with variance σ^2. In this section, we explain how to adapt the fidelity term to deal with the statistics of SAR images.

As mentioned in Sect. 4.1.3, under Goodman's model, noise fluctuations in SAR intensity images are multiplicative and Gamma distributed. A classical approach, known as the *homomorphic* transform, is thus to consider the log transform of the image in order to turn multiplicative fluctuations into additive ones as already stated in Sects. 4.1.3 and 4.2.3.

As a consequence, the problem becomes similar to the Gaussian case, such that the MAP problem can be expressed with $\ell(\tilde{v}_i|\tilde{u}_i) = \frac{1}{2\sigma^2}(\tilde{v}_i - \tilde{u}_i)^2$ where $\tilde{v}_i = \log v_i$

and $\tilde{u}_i = \log u_i$. The image of interest $\mathbf{u}$ can be obtained from $\tilde{\mathbf{u}}$ by taking the exponential after the optimization is performed.

As already mentioned in the Sect. 4.2.3, bias correction should be applied, leading to the inversion formula: $u_i = \exp(\tilde{u}_i + \log L - \Psi(L))$. This strategy has been extensively used, e.g., for wavelet prior in [95], for patch-based priors as in KSVD [96] or for non-local filtering [97].

Instead of reducing the original problem to the Gaussian case, one can directly consider the neg-log-likelihood function of Gamma random variables defined, up to an additive constant, by (see box in Sect. 4.1.3.4):

$$\ell(v_i|u_i) = L \log u_i + L\frac{v_i}{u_i} \tag{4.72}$$

where L is the number of looks. However, unlike a square fidelity term, this neg-log-likelihood term is a non-convex function of u_i as illustrated in Fig. 4.5. As a consequence, the subsequent optimization problems are more difficult to solve, all the more since several patch-based priors previously described have a non-convex energy. In order to alleviate this issue, several authors [64, 66] have suggested to optimize not with respect to u_i but by considering again its log transform $\log u_i$. In fact, using the changes of variable $\tilde{u}_i = \log u_i$ and $\tilde{v}_i = \log v_i$, we obtain:

$$\ell(v_i|u_i) = \tilde{\ell}(\tilde{v}_i|\tilde{u}_i) = L\tilde{u}_i + L\exp(\tilde{v}_i - \tilde{u}_i) \tag{4.73}$$

and we can easily check that this function is convex with respect to $\tilde{u}_i$. Remark that $\tilde{\ell}$ is, up to an additive constant, the neg-log-likelihood of the Fisher–Tippett distribution modeling the distribution of log transformed Gamma random variables. Unlike the homomorphic approach, this procedure is free of approximation and then free of systematic bias under Goodman's assumptions, such that one can take the exponential after the optimization is performed: $u_i = \exp(\tilde{u}_i)$. This was the approach followed for adapting EPLL to SAR intensities in [94] and illustrated in Fig. 4.9.

It is important to keep in mind that such change of variable is usually applied only on the likelihood term but not on the prior term, which subsequently changes the optimization problem and the nature of the solution. While this technique simplifies the optimization procedure, it changes the prior now expressed on the log transform of the sought solution.

4.3.1.3 Parameter Tuning

Estimation of the parameters of interest $\mathbf{u}$ by maximization of the posterior probability (i.e., MAP estimation) depends on the tuning of several additional parameters. The first parameter is the relative weight of the regularization (prior term) with respect to the data fidelity term (likelihood). Patch-based priors also depend on the choice of the size of patches, the size K of the dictionary or number K of components of the Gaussian mixture, and more importantly, on the parameter β in the variable splitting for the alternate minimization strategy (see Eqs. (4.67) and (4.71)).

Patch-based MRF models capture complex textures and structures, thereby providing high-quality restorations with limited blurring of sharp features or textured areas. These methods, however, require an heavy learning step (model learning typically requires a day—up to several days—on a single core) and the optimization step to estimate an image once the model is learnt is also challenging (non-convex, high-dimensional). Extension to multivariate SAR modalities (i.e., polarimetric, interferometric) is non-trivial.

4.3.2 *Patch-Based Selection for Estimation in Polarimetric or Interferometric SAR*

With the polarimetric and interferometric modalities, each pixel of a SAR image gathers measurements from several polarimetric or interferometric channels. Modeling the statistical dependence between channels and between pixels within a patch is challenging because of the increased dimension and nonlinear relationship between parameters of interest (radiometry, interferometric phase, polarimetric properties) and the SAR measurements. Rather than defining a prior that models the complex relationship within patches, an alternative still based on patches is to exploit patch redundancy to *select* robustly similar pixels. Selection-based estimation has been proposed long ago by Lee [98] and later extended by comparing patches to improve the robustness of the selection by Buades et al. [99]. This latter approach is known as the *non-local means* since selected pixels can potentially be far apart without being necessarily connected.

Generalized likelihood ratio in SAR imagery:

Under Goodman's model (see Sect. 4.1.3), the GLR between two (multivariate) pixel values v_1 and v_2 is given below for the different considered SAR modalities with number of looks L. When GLR is used on patches, these quantities have to be summed over the pairs of corresponding pixels in both patches.

- **SAR Amplitude:** $$\delta(\mathrm{A}_1, \mathrm{A}_2) = 2L \log\left[\frac{1}{2}\left(\frac{\mathrm{A}_1}{\mathrm{A}_2} + \frac{\mathrm{A}_2}{\mathrm{A}_1}\right)\right].$$
- **SAR Intensity:** $$\delta(\mathrm{I}_1, \mathrm{I}_2) = 2L \log\left[\frac{1}{2}\left(\sqrt{\frac{\mathrm{I}_1}{\mathrm{I}_2}} + \sqrt{\frac{\mathrm{I}_2}{\mathrm{I}_1}}\right)\right].$$
- **InSAR** ($L = 1$)**:** $$\delta(\mathrm{Z}_1, \mathrm{Z}_2) = 2 \log\left[\frac{(|\mathrm{Z}_1|^2 + |\mathrm{Z}'_1|^2 + |\mathrm{Z}_2|^2 + |\mathrm{Z}'_2|^2)^2 - 4|\mathrm{Z}_1^* \mathrm{Z}'_1 + \mathrm{Z}_2^* \mathrm{Z}'_2|^2}{|(|\mathrm{Z}_1|^2 - |\mathrm{Z}'_1|^2)(|\mathrm{Z}_2|^2 - |\mathrm{Z}'_2|^2)|}\right].$$

- **Scattering Vector** $(L = 1)$**:** undefined, use instead the next criterion with $\mathbf{C} = \mathbf{k}\mathbf{k}^{*t}$.
- **SAR Covariance Matrix:**

  if $L < D$, use the next criterion with a full-rank approximation of $\mathbf{C}$.

  otherwise, $\delta(\mathbf{C}_1, \mathbf{C}_2) = 2L \log\left(\dfrac{\frac{1}{2}|\mathbf{C}_1 + \mathbf{C}_2|}{\sqrt{|\mathbf{C}_1||\mathbf{C}_2|}}\right)$.

4.3.2.1 Patch Similarity

The estimators considered here rely on the detection of a large number of pixels j sharing the same underlying physical parameters as pixel i, i.e., $u_j = u_i$. In practice $\mathbf{u}$ is unknown, and this information should be measured by a robust dissimilarity criterion depending only on the input noisy image $\mathbf{v}$.

Choice of the dissimilarity measure. In light of detection theory, deciding that two noisy observations v_1 and v_2 share the same noise-free value can be performed by confronting a null-hypothesis $\mathscr{H}_0 : u_1 = u_2$ against an alternative one $\mathscr{H}_1 : u_1 \neq u_2$. The generalized likelihood ratio (GLR) is designed for this hypothesis test [100, 101] and thus can be used to define a dissimilarity criterion as:

$$\delta(v_1, v_2) = -\log \frac{\sup_{u_1} \mathrm{p}(v_1|u_1) \sup_{u_2} \mathrm{p}(v_2|u_2)}{\sup_{u_{12}} \mathrm{p}(v_1|u_{12})\mathrm{p}(v_2|u_{12})} \tag{4.74}$$

In many situations of interest, GLR has a closed form derived directly from the expression of the ML estimate (see the box "Generalized likelihood ratio in SAR imagery" on p. 177). However, the GLR is not guaranteed to exist, for instance, in multivariate SAR imagery with $L < D$. Performing diagonal loading of the ML estimates before evaluating the GLR is a practical alternative explored in [102].

The GLR provides a method to measure similarities in a way that is adapted to the distribution of noise; its performance is, however, limited for high noise levels since a large dissimilarity can be equally ascribed to the noise or to an intrinsic difference.

From pixels to patches. Pixel-based comparisons being sensitive to the noise, the authors of [99] have suggested exploiting the redundancy of patches by comparing patches instead of pixels:

$$\Delta(\mathbf{v}_{\square i}, \mathbf{v}_{\square j}) = \frac{1}{m} \sum_{k=1}^{m} \delta\big((\mathbf{v}_{\square i})_k,\ (\mathbf{v}_{\square j})_k\big) \tag{4.75}$$

where the notation $(\mathbf{v}_{\square i})_k$ stands for the kth pixel of the patch centered at pixel i of image $\mathbf{v}$. In general, the variance of $\Delta(\mathbf{v}_{\square i}, \mathbf{v}_{\square j})$ is m times smaller than that of $\delta(v_i, v_j)$. As a consequence, patch comparison is much more robust to noise and large dissimilarities can reasonably be ascribed to intrinsic differences. Nevertheless, patches should not be too large; otherwise, most of the pairs of patches $\mathbf{v}_{\square i}$ and $\mathbf{v}_{\square j}$ will be dissimilar. The patch size should thus be chosen wisely regarding both the noise level and the image content in order to reach a good trade-off in terms of detection error.

Improving discrimination versus localization. There is an alternative to using larger patches for reducing the variance of patch similarity Δ: pre-filtering (i.e., spatial averaging) the image before performing patch comparisons. This way, the size of the patches can be kept small enough so that many similar patches can be found in an extended neighborhood while limiting the variance of the dissimilarity, hence preserving its selectivity. Such a pre-filtering improves the detection of low contrasted structures at a price of a loss of its localization. The strength of the pre-filtering should thus be again chosen wisely regarding both the noise level and the image content.

4.3.2.2 Local Maximum Likelihood-Based Estimators

By construction patch similarity provides a measure of how much a patch $\mathbf{v}_{\square j}$ is identically distributed with $\mathbf{v}_{\square i}$, in particular it offers a robust statistical test for $u_j = u_i$. From this measure, we can thus define a soft-assignment $\pi_{i,j} = \varphi(\Delta(\mathbf{v}_{\square i}, \mathbf{v}_{\square j}))$, where $\varphi : \mathbb{R}^+ \to [0, 1]$ is a kernel function chosen such that $\pi_{i,j}$ gets closer to 1 when $u_i = u_j$ becomes more likely. The kernel is often chosen as $\varphi(\cdot) = \exp(-\cdot/h)$, for $h > 0$, but other choices can be used [102]. These assignments provide a key ingredient to artificially increase the number of observations available at each pixel location i. While prior regularization is necessary when only one observation is available at each pixel i, it can be avoided providing enough pixels are selected.

Weighted maximum likelihood in SAR imagery:

Under Goodman's model (see Sect. 4.1.3). the WML estimator reads for the different considered SAR modalities, with number of looks L, as

- **SAR Amplitude:** $\widehat{\mathrm{R}}_i^{(\mathrm{WML})} = \sum_j w_{i,j} \mathrm{A}_j^2 .$
- **SAR Intensity:** $\widehat{\mathrm{R}}_i^{(\mathrm{WML})} = \sum_j w_{i,j} \mathrm{I}_j .$

- **InSAR** $(L = 1)$**:**
$$\begin{cases} \widehat{\mathrm{R}}_i^{(\mathrm{WML})} = \frac{1}{2}\Big(\sum_j w_{i,j}|\mathrm{Z}_j|^2 + \sum_j w_{i,j}|\mathrm{Z}'_j|^2\Big), \\ \widehat{\varphi}_i^{(\mathrm{WML})} = \arg \sum_j w_{i,j}\mathrm{Z}_j^*\mathrm{Z}'_j, \\ \widehat{\rho}_i^{(\mathrm{WML})} = \left| \dfrac{2\sum_j w_{i,j}\mathrm{Z}_j^*\mathrm{Z}'_j}{\sum_j w_{i,j}|\mathrm{Z}_j|^2 + \sum_j w_{i,j}|\mathrm{Z}'_j|^2} \right|. \end{cases}$$
- **Scattering Vector** $(L = 1)$**:** $\widehat{\mathbf{\Sigma}}_i^{(\mathrm{WML})} = \sum_j w_{i,j}\mathbf{k}_j\mathbf{k}_j^{*t}$.
- **SAR Covariance Matrix:** $\widehat{\mathbf{\Sigma}}_i^{(\mathrm{WML})} = \sum_j w_{i,j}\mathbf{C}_j$, with $w_{i,j} = \dfrac{\pi_{i,j}}{\sum_j \pi_{i,j}}$.

Weighted maximum likelihood: The weighted maximum likelihood (WML) estimator, introduced in [103], uses the assignment $\pi_{i,j} \in [0, 1]$, representing the confidence that pixels j are identical to pixels i, in order to estimate $\mathbf{u}$ as:

$$\widehat{\mathbf{u}}^{(\mathrm{WML})} \in \arg\min_{\mathbf{u}} \sum_i \sum_j \pi_{i,j}\, \ell(v_j|u_i) \tag{4.76}$$

The above optimization problem is clearly separable in i and the solution $\widehat{u}_i^{(\mathrm{WML})}$ is often known in closed form (see the box "Weighted maximum likelihood in SAR imagery" on p. 43). As soon as all assigned samples are identically distributed, this estimator is unbiased meaning that it provides in average the sought physical parameters $\mathbb{E}[\widehat{\mathbf{u}}^{(\mathrm{WML})}] = \mathbf{u}$. The WML estimator is thus able to estimate $\mathbf{u}$, without prior regularization, providing a large number of independent and identically distributed samples j is assigned to i. Indeed, in many situations, the WML estimation consists in performing a weighted average of the selected pixels, for which the noise variance is reduced at pixel index i by a factor:

$$\widehat{L}_i^{(\mathrm{WML})} = \frac{\mathrm{Var}[v_i]}{\mathrm{Var}[\widehat{u}_i^{(\mathrm{WML})}]} = \frac{(\sum \pi_{i,j})^2}{\sum \pi_{i,j}^2} \tag{4.77}$$

provided that the assigned samples are independent. For a hard assignment, i.e., $\pi_{i,j} \in \{0, 1\}$, this quantity is exactly the number of selected pixels and this estimator coincides with the ML estimator known to be efficient, i.e., its variance reaches the so-called Cramer–Rao lower bound (see, e.g., [104]). Nevertheless, rather than following an "all or nothing" selection strategy, the authors of [103] showed that it is often safer to use real-valued weights $\pi_{i,j} \in [0, 1]$. The soft-assignments $\pi_{i,j}$ provide more flexibility by reducing more the noise variance at a price of a larger bias introduced by heterogeneous samples. It can thus be important to perform a weight correction to limit such bias.

Reweighted maximum likelihood: An heterogeneous collection of samples presents larger variations than those due only to the noise fluctuations. In SAR imaging,

the noise is signal-dependent so that the noise variance can be predicted given the underlying parameter u_i by a function $\sigma^2 : u_i \rightarrow \mathrm{Var}[v_i]$, e.g., $\sigma^2(u_i) = u_i^2/L$ for the Gamma distribution. By comparing the predicted variance $\sigma^2(\widehat{u_i}^{(\mathrm{WML})})$ with the empirical variance of the (soft-) assigned samples:

$$\widehat{\sigma}_i^{2(\mathrm{WML})} = \sum_j w_{i,j} v_j^2 - \Big(\sum_j w_{i,j} v_j\Big)^2 \tag{4.78}$$

it can be checked if identical pixels have been selected. A mismatch indicates that candidates belong to different populations and the WML estimation may be biased. Inspired from [105], weights can be reevaluated as:

$$\tilde{\pi}_{i,j} = \begin{cases} (1-\alpha_i)\pi_{i,i} + \alpha_i(\sum_j \pi_{i,j}) & \text{if} \quad i = j \\ (1-\alpha_i)\pi_{i,j} & \text{otherwise} \end{cases} \tag{4.79}$$

$$\text{with} \quad \alpha_i = \max\left[\frac{\widehat{\sigma}_i^{2(\mathrm{WML})} - \sigma^2(\widehat{u_i}^{(\mathrm{WML})})}{\sigma^2(\widehat{u_i}^{(\mathrm{WML})})}, 0\right] \tag{4.80}$$

The original weights are thus kept when there is a perfect match between predicted and empirical variances ($\alpha_i = 0$), and their mass is redistributed on the pixel index i when they completely mismatch ($\alpha_i = 1$). Provided that the WML estimation coincides with the weighted average, the reweighted solution simply reads as $\widehat{u}_i^{(\mathrm{RWML})} = (1-\alpha_i)\widehat{u}_i^{(\mathrm{WML})} + \alpha_i v_i$, and the amount of noise reduction $\widehat{L}_i^{(\mathrm{RWML})}$ can easily be updated from $\widehat{L}_i^{(\mathrm{WML})}$ [102]. The RWML estimator trades a reduction of the bias of the WML estimator for an increase of residual variance. It follows that the RWML estimator is nearly unbiased and the amount of noise reduction $\widehat{L}_i^{(\mathrm{RWML})}$ becomes a local measure of its performance.

Adaptive reweighted maximum likelihood. The quality of the weights $\pi_{i,j}$ depends on several parameters: the patch size, the pre-filtering strength but also the kernel function φ and the size of the search window inside which the patches $\mathbf{v}_{\square j}$ are extracted. The NL-SAR algorithm of [102] proposed a practical way to compute efficiently results obtained with a hundred of different combinations of parameters. For each of them, the measure of quality $\widehat{L}_i^{(\mathrm{RWML})}$ is used to decide at each pixel i which solution should be preferred. The resulting estimator adapts its own internal parameters automatically to the local content of the image. This has shown to be a robust solution to deal with the high variability of SAR scenes, especially in multivariate contexts, and without requiring any manual parameter tuning. Figure 4.10 gives an illustration of NL-SAR on an airborne polarimetric SAR image.

4.3.2.3 Selection-Based Collaborative Filtering

Unlike WML estimators using patch similarity to artificially increase the number of local observations, patch similarity can be used to define a prior model promoting the regularity of the stacks of similar patches.

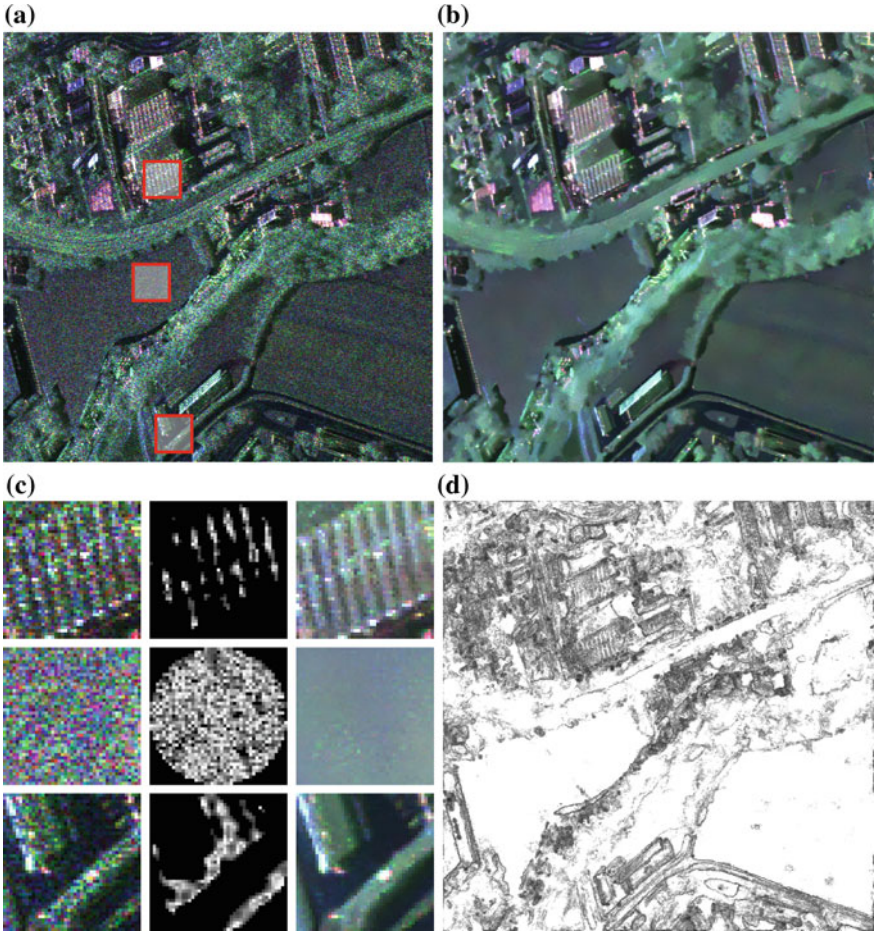


Fig. 4.10 **a** Original single-look polarimetric FSAR image (© DLR) of Kaufbeuren (Germany). The three-dimensional complex scattering vectors $\mathbf{k}_i$ are represented using an RGB representation from the Pauli basis. **b** Estimation of the covariance matrices $\widehat{\boldsymbol{\Sigma}}_i$ by NL-SAR [94], **c** From *left* to *right*, three different windows located as highlighted in **a**, their corresponding weights (range [0, 1]) used to estimate the central pixel, and the final denoising result of these windows. Remark that the selection adapts to the image content and the selected pixels do not form a connected component, thus referred to as a *non-local* neighborhood. **d** Map of the equivalent number of looks $\widehat{L}_i^{(\mathrm{RWML})}$ encoding the local amount of noise reduction (range [0, 100])

While a set of heterogeneous patches is difficult to model a priori, all the more when the image is multivariate, a stack of similar patches has a high redundancy and is thus much easier to model and regularize. In a variational stack-based approach, the prior distribution of $\mathbf{u}$ can be expressed in terms of the prior distribution of all its stacks of patches. Since patch similarity is measured on $\mathbf{v}$, this prior is taken conditionally to the observation $\mathbf{v}$ (in the same vein as for CRFs) leading to:

$$\mathrm{p}(\mathbf{u}|\mathbf{v}) = \frac{1}{\mathscr{Z}} \exp\Big\{-\sum_{p\in\mathscr{P}} \theta_s\big(\mathbf{u}_{\square p}\big)\Big\}\,, \tag{4.81}$$

where $\mathscr{Z}$ is the partition function (see Eq. (4.41)), $\mathbf{u}_{\square p} = (\mathbf{u}_{\square q})$ is a stack of patches of $\mathbf{u}$ whose corresponding patches in $\mathbf{v}$ (or some pre-filtered version of it) are similar, e.g., such that $\Delta(\mathbf{v}_{\square q}, \mathbf{v}_{\square p}) \leq \tau$ for some $\tau > 0$ (usually the patches of the stacks are sorted according to Δ and similar patches are sought within an extended search area that is much smaller than the whole image domain) and the functions $\theta_s(\cdot)$ are the *stack energies*. The cube in the notation $\mathbf{u}_{\square p}$ is used to emphasize that the stack is a three-dimensional signal with fewer pixels than the dimension n of the whole image $\mathbf{u}$. The MAP estimation can thus be formulated as:

$$\widehat{\mathbf{u}}^{(\mathrm{MAP})} \in \arg\min_{\mathbf{u}} \sum_i \ell(v_j|u_j) + \sum_{p\in\mathscr{P}} \theta_s(\mathbf{u}_{\square p})\,. \tag{4.82}$$

This problem is difficult to tackle in the present form, and many studies have considered instead performing a collaborative filtering (CF). CF consists of computing the solutions (i.e., estimates) of K smaller and independent optimization subproblems expressed on stacks, and merging these solutions to reconstruct an image so that all its stacks are in good agreement with the restored ones, for instance, as:

$$\widehat{\mathbf{u}}^{(\mathrm{CF})} = \arg\min_{\mathbf{u}} \sum_{p\in\mathscr{P}} \big\| \widehat{\mathbf{u}}^{(\mathrm{MAP})}_{\square p} - \mathbf{u}_{\square p} \big\|^2\,, \tag{4.83}$$

$$\text{where}\quad \widehat{\mathbf{u}}^{(\mathrm{MAP})}_{\square p} = \arg\min_{\mathbf{u}_{\square}} \sum_k \ell((\mathbf{v}_{\square p})_k|(\mathbf{u}_{\square})_k) + \theta_s(\mathbf{u}_{\square})\,, \tag{4.84}$$

with $\mathbf{v}_{\square p}$ the noisy stack formed by collecting noisy patches that are located at the same position as those of $\mathbf{u}_{\square p}$.

The block matching 3D (BM3D) [106] and the non-local Bayes [107] are particular examples of collaborative filters obtained from different stack energies: the first one enforces regularity in a Fourier or wavelet domain, while the second uses a Gaussian prior adapted to each stack. In these algorithms, CF is performed in two steps by first defining stacks from $\mathbf{v}$ and next reupdating the stacks from the similarity of patches in $\widehat{\mathbf{u}}^{(\mathrm{CF})}$ obtained at the first step. The SAR-BM3D algorithm [108] is an extension of BM3D with a data fidelity term suitable for the intensity of SAR images. The extension of CF to multivariate SAR data still remains an open problem.

4.4 Conclusion

In this chapter, we have presented two powerful mathematical frameworks, Markov random fields and patch-based approaches, to process very-high-resolution SAR imagery. Although these models can be used for medium-resolution SAR images,

they are specially adapted to high resolution since they are able to take into account fine textural patterns in the images. We have seen that both models can be exploited for different SAR modalities (amplitude, interferometry, polarimetry, multi-channel SAR) thanks to the taking into account of the statistics of the SAR imagery. Both models rely on different assumptions (local smoothness and local redundancy, respectively) that can be broken into different areas. For instance, some patches may be very rare in the SAR images, and the signals can be locally discontinuous, both situations appearing specially in the surrounding of very bright scatterers. In this case, the combination of both models may improve the results by taking the best part of both of them. This is the case, for instance, in [109] for multi-channel 3D InSAR reconstruction.

Further Readings

Concerning SAR acquisition systems and radar data synthesis, the following books give detailed presentations [110, 111], the second focusing mainly on polarimetry.

Markov random fields have been the subject of many developments for more than 30 years. The book [33] provides a general overview on the topic while the survey [34] points recent advances. Optimization methods suited to inference problems in Markov random fields are compared in [45] on several representative cases. The exploitation of MRFs in the framework of multi-channel phase unwrapping is described and analyzed in [77].

The interested reader will find in the two review papers [112, 113] a detailed presentation of speckle reduction methods for SAR imagery.

References

1. Moreira, A., Prats-Iraola, P., Younis, M., Krieger, G., Hajnsek, I., Papathanassiou, K.P.: A tutorial on synthetic aperture radar. IEEE Trans. Geosci. Remote Sens. **1**(1), 6–43 (2013)
2. Fornaro, G., Pascazio, V.: Chapter 20 - SAR interferometry and tomography: theory and applications. In: Academic Press Library in Signal Processing: Communications and Radar Signal Processing. Academic Press Library in Signal Processing, vol. 2, pp. 1043–1117. Elsevier (2014)
3. Franceschetti, G., Schirinzi, G.: A SAR processor based on two-dimensional FFT codes. IEEE Trans. Aerosp. Electron. Syst. **26**(2), 356–366 (1990)
4. Franceschetti, G., Lanari, R.: Synthetic aperture radar processing (1999)
5. Franceschetti, G., Lanari, R., Pascazio, V., Schirinzi, G.: WASAR: a wide-angle SAR processor. IEE Proc. Part F: Radar Signal Process. **139**(2), 107–114 (1992)
6. Bamler, R., Hartl, P.: Synthetic Aperture radar interferometry. Inverse Problem **14**, R1–R54 (1998)
7. Bao, M., Bruning, C., Alpers, W.: Simulation of ocean waves imaging by an along-track interferometric synthetic aperture radar. IEEE Trans. Geosci. Remote Sens. **35**(3), 618–631 (1997)
8. Budillon, A., Pascazio, V., Schirinzi, G.: Estimation of radial velocity of moving targets by along-track interferometric SAR systems. IEEE Geosci. Remote Sens. Lett. **5**(3), 349–353 (2008)
9. Goldstein, R.M., Zebker, H.A., Werner, C.L.: Satellite radar interferometry: two-dimensional phase unwrapping. Radio Sci. **31**, 445–464 (1993)

10. Tupin, F., Inglada, J., Nicolas, J.M.: Remote Sensing Imagery. ISTE - Wiley (2014)
11. Goodman, J.: Some fundamental properties of speckle. J. Opt. Soc. Am. **66**(11), 1145–1150 (1976)
12. Abramowitz, M., Stegun, I.: Handbook of Mathematical Functions. Dover Publications (1964)
13. Lee, J., Hoppel, K., Mango, S., Miller, A.: Intensity and phase statistics of multilook polarimetric and interferometric SAR imagery. IEEE Trans. Geosci. Remote Sens. **32**(5), 1017–1028 (1994)
14. Nicolas, J.M.: Introduction aux statistiques de deuxème espèce: applications des log-moments et des log-cumulants à l' analyse des lois d'image radar. Traitement du signal **19**(3), 139–167 (2002)
15. Frery, A.C., Muller, H.J., Yanasse, C.C.F., Sant'Anna, S.J.S.: A model for extremely heterogeneous clutter. IEEE Trans. Geosci. Remote Sens. **35**(3), 648–659 (1997)
16. Krylov, V., Moser, G., Serpico, S.B., Zerubia, J.: On the method of logarithmic cumulants for parametric probability density function estimation. IEEE Trans. Image Process. **22**(10), 3791–3806 (2013)
17. Tison, C., Nicolas, J.M., Tupin, F., Maître, H.: A new statistical model of urban areas in high resolution SAR images for Markovian segmentation. IEEE Trans. Geosci. Remote Sens. **42**, 2046–2057 (2004)
18. Anfinsen, S.N., Eltoft, T.: Application of the matrix-variate Mellin transform to analysis of polarimetric radar images. IEEE Trans. Geosci. Remote Sens. **49**(6), 2281–2295 (2011)
19. Nicolas, J.M., Tupin, F.: Statistical models for SAR amplitude data: a unified vision through Mellin transform and Meijer functions. In: EUSIPCO. Budapest, Hungary (2016)
20. Besag, J.: Spatial interaction and the statistical analysis of lattice systems. J. R. Stat. Soc. Ser. B (Methodol.), pp. 192–236 (1974)
21. Lafferty, J., McCallum, A., Pereira, F.C.: Conditional Random Fields: Probabilistic models for segmenting and labeling sequence data (2001)
22. Fjorthoft, R., Delignon, Y., Pieczinski, W., Sigelle, M., Tupin, F.: Unsupervised segmentation of radar images using hidden Markov chains and hidden Markov random fields. IEEE Trans. Geosci. Remote Sens. pp. 675–686 (2003)
23. Benboudjema, D., Tupin, F.: Markovian modeling and Fisher distribution for unsupervised segmentation of radar images. Int. J. Remote Sens. (2013)
24. Niu, X., Ban, Y.: Unsupervised SAR image segmentation using a hierarchical TMF model. IEEE Geosci. Remote Sens. Lett. **7**(1), 210–214 (2010)
25. Voisin, A., Krylov, V., Moser, G., Serpico, S.B., Zerubia, J.: Classification of very high resolution SAR images of urban areas using copulas and texture in a hierarchical Markov random field model. IEEE Geosci. Remote Sens. Lett. **10**(1), 96–100 (2013)
26. Wang, Y.H., Han, C.Z., Tupin, F.: PolSAR data segmentation by combining tensor space cluster analysis and Markovian framework. IEEE Geosci. Remote Sens. Lett. **10**(5) (2013)
27. Lobry, S., Denis, L., Tupin, F.: Multitemporal SAR image decomposition into strong scatterers, background, and speckle. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. (2016)
28. Niu, X., Ban, Y.: An Adaptive Contextual SEM Algorithm for Urban Land Cover Mapping Using Multitemporal High-Resolution Polarimetric SAR Data. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. **5**(4) (2012)
29. Moser, G., Serpico, S.B.: Unsupervised change detection from multichannel SAR data by Markovian data fusion. IEEE Trans. Geosci. Remote Sens. **47**(7), 2114–2128 (2009)
30. Perciano, T., Tupin, F., Hirata, R., Cesar, R.M.: A two-level Markov random field for road network extraction and its application with optical, SAR and multitemporal data. Int. J. Remote Sens. (2016)
31. Benedek, C., Descombes, X., Zerubia, J.: Building development monitoring in multitemporal remotely sensed image pairs with stochastic birth-death dynamics. IEEE Trans. Pattern Anal. Mach. Intell. **34**(1), 33–50 (2012)
32. Neal, R.M.: Probabilistic inference using Markov chain Monte Carlo methods (1993)
33. Li, S.Z.: Markov Random Field Modeling in Image Analysis. Springer Science & Business Media (2009)

34. Markov Random Field modeling: inference & learning in computer vision & image understanding: a survey. Comput. Vis. Image Underst. **117**(11), 1610–1627 (2013)
35. Pock, T., Cremers, D., Bischof, H., Chambolle, A.: Global solutions of variational models with convex regularization. SIAM J. Imaging Sci. **3**(4), 1122–1145 (2010)
36. Nocedal, J., Wright, S.: Numerical Optimization. Springer Science & Business Media (2006)
37. Liu, D.C., Nocedal, J.: On the limited memory BFGS method for large scale optimization. Math. Program. **45**(1–3), 503–528 (1989)
38. Combettes, P.L., Pesquet, J.C.: Proximal splitting methods in signal processing. In: Fixed-Point Algorithms for Inverse Problems in Science and Engineering, pp. 185–212. Springer (2011)
39. Parikh, N., Boyd, S.P.: Proximal algorithms. Found. Trends Opt. **1**(3), 127–239 (2014)
40. Bertsekas, D.P.: Constrained Optimization and Lagrange Multiplier Methods. Academic press (2014)
41. Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J.: Distributed optimization and statistical learning via the alternating direction method of multipliers. Found. Trends® Mach. Learn. **3**(1), 1–122 (2011)
42. Hestenes, M.R.: Multiplier and gradient methods. J. Opt. Theory Appl. **4**(5), 303–320 (1969)
43. Pearl, J.: Probabilistic reasoning in intelligent systems: networks of plausible inference (1988)
44. Freeman, W.T., Pasztor, E.C., Carmichael, O.T.: Learning low-level vision. Int. J. Comput. Vis. **40**(1), 25–47 (2000)
45. Kappes, J.H., Andres, B., Hamprecht, F.A., Schnörr, C., Nowozin, S., Batra, D., Kim, S., Kausler, B.X., Kröger, T., Lellmann, J., et al.: A comparative study of modern inference techniques for structured discrete energy minimization problems. Int. J. Comput. Vis. **115**(2), 155–184 (2015)
46. Greig, D.M., Porteous, B.T., Seheult, A.H.: Exact maximum a posteriori estimation for binary images. J. R. Statist. Soc. B **51**(2), 271–279 (1989)
47. Ford, L.R., Fulkerson, D.R.: Maximal flow through a network. Can. J. Math. **8**(3), 399–404 (1956)
48. Cormen, T.H., Leiserson, C.E., Rivest, R.L., Stein, C.: Introduction to Algorithms. MIT Press, McGraw-Hill Book Co. (2001)
49. Boykov, Y., Kolmogorov, V.: An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. IEEE Trans. Pattern Anal. Mach. Intell. **26**(9), 1124–1137 (2004)
50. Verma, T., Batra, D.: MaxFlow revisited: an empirical comparison of maxflow algorithms for dense vision problems. In: BMVC, pp. 1–12 (2012)
51. Kolmogorov, V.: `maxflow`, a C++ library for maxflow/min-cut computation (2010). http://vision.csd.uwo.ca/code/. Accessed June 2016
52. Kolmogorov, V., Zabih, R.: What energy functions can be minimized via graph-cuts? IEEE Trans. Pattern Anal. Mach. Intell. **26**(2) (2004)
53. Kolmogorov, V., Rother, C.: Minimizing nonsubmodular functions with graph cuts-a review. IEEE Trans. Pattern Anal. Mach. Intell. **29**(7), 1274–1279 (2007)
54. Boykov, Y., Veksler, O., Zabih, R.: Fast approximate energy minimization via graph cuts. IEEE Trans. Pattern Anal. Mach. Intell. **23**(11), 1222–1239 (2001)
55. Lempitsky, V., Rother, C., Roth, S., Blake, A.: Fusion moves for Markov random field optimization. IEEE Trans. Pattern Anal. Mach. Intell. **32**(8), 1392–1405 (2010)
56. Lempitsky, V., Rother, C., Blake, A.: Logcut-efficient graph cut optimization for Markov random fields. In: IEEE International Conference on Computer Vision, pp. 1–8 (2007)
57. Ishikawa, H.: Exact optimization for Markov random fields with convex priors. IEEE Trans. Pattern Anal. Mach. Intell. **25**(10), 1333–1336 (2003)
58. Darbon, J., Sigelle, M.: Image restoration with discrete constrained total variation part ii: Levelable functions, convex priors and non-convex cases. J. Math. Imaging Vis. **26**(3), 277–291 (2006)
59. Schlesinger, D., Flach, B.: Transforming an arbitrary minsum problem into a binary one. Fak. Informatik, TU (2006)

60. Aubert, G., Aujol, J.F.: A variational approach to removing multiplicative noise. SIAM J. Appl. Math. **68**(4), 925–946 (2008)
61. Denis, L., Tupin, F., Darbon, J., Sigelle, M.: SAR image regularization with fast approximate discrete minimization. IEEE Trans. Image Process. **18**(7), 1588–1600 (2009)
62. Rudin, L.I., Osher, S., Fatemi, E.: Nonlinear total variation based noise removal algorithms. Phys. D: Nonlinear Phenom. **60**(1), 259–268 (1992)
63. Xie, H., Pierce, L.E., Ulaby, F.T.: Statistical properties of logarithmically transformed speckle. IEEE Trans. Geosci. Remote Sens. **40**(3), 721–727 (2002)
64. Bioucas-Dias, J.M., Figueiredo, M.A.: Multiplicative noise removal using variable splitting and constrained optimization. IEEE Trans. Image Process. **19**(7), 1720–1730 (2010)
65. Durand, S., Fadili, J., Nikolova, M.: Multiplicative noise removal using L1 fidelity on frame coefficients. J. Math. Imaging Vis. **36**(3), 201–226 (2010)
66. Steidl, G., Teuber, T.: Removing multiplicative noise by Douglas-Rachford splitting methods. J. Math. Imaging Vis. **36**(2), 168–184 (2010)
67. Çetin, M., Karl, W.C.: Feature-enhanced synthetic aperture radar image formation based on nonquadratic regularization. IEEE Trans. Image Process. **10**(4), 623–631 (2001)
68. Potter, L.C., Ertin, E., Parker, J.T., Cetin, M.: Sparsity and compressed sensing in radar imaging. Proc. IEEE **98**(6), 1006–1020 (2010)
69. Denis, L., Tupin, F., Rondeau, X.: Exact discrete minimization for TV + L0 image decomposition models. In: IEEE International Conference on Image Processing, pp. 2525–2528 (2010)
70. Denis, L., Tupin, F., Darbon, J., Sigelle, M.: Joint regularization of phase and amplitude of InSAR data: Application to 3-D reconstruction. IEEE Trans. Geosci. Remote Sens. **47**(11), 3774–3785 (2009)
71. Itoh, K.: Analysis of the phase unwrapping problem. Appl. Opt. **21**(14) (1982)
72. Eineder, M., Adam, N.: A maximum-likelihood estimator to simultaneously unwrap, geocode, and fuse SAR interferograms from different viewing geometries into one digital elevation model. Appl. Numer. Math. **43**(4), 359–373 (2002)
73. Ferretti, A., Prati, C., Rocca, F.: Multibaseline InSAR DEM reconstruction: the wavelet approach. Il Nuovo Cimento **24**, 159–176 (2001)
74. Fornaro, G., Pauciullo, A., Sansosti, E.: Phase difference-based multichannel phase unwrapping. IEEE Trans. Image Process. **14**(7), 960–972 (2005)
75. Pascazio, V., Schirinzi, G.: Multifrequency InSAR height reconstruction through maximum likelihood estimation of local planes parameters. IEEE Trans. Image Process. **11**(12), 1478–1489 (2002)
76. Ferraiuolo, G., Meglio, F., Pascazio, V., Schirinzi, G.: DEM reconstruction accuracy in multichannel SAR interferometry. IEEE Trans. Geosci. Remote Sens. **47**(1), 191–201 (2009)
77. Baselice, F., Ferraioli, G., Pascazio, V., Schirinzi, G.: Contextual information-based multichannel synthetic aperture radar interferometry: addressing dem reconstruction using contextual information. IEEE Signal Process. Mag. **31**(4), 59–68 (2014)
78. Ferraiuolo, G., Pascazio, V., Schirinzi, G.: Maximum a posteriori estimation of height profiles in InSAR imaging. IEEE Geosci. Remote Sens. Lett. **1**, 66–70 (2004)
79. Ferraioli, G., Shabou, A., Tupin, F., Pascazio, V.: Multichannel phase unwrapping with graph-cuts. IEEE Geosci. Remote Sens. Lett. **6**(3), 562–566 (2009)
80. Ferraioli, G., Ferraiuolo, G., Pascazio, V.: Phase-Offset Estimation in Multichannel SAR Interferometry. IEEE Geosci. Remote Sens. Lett. **5**(3), 458–462 (2008)
81. Gatti, G., Tebaldini, S., Mariotti d'Alessandro, M., Rocca, F.: ALGAE: a fast algebraic estimation of interferogram phase offsets in space-varying geometries. IEEE Trans. Geosci. Remote Sens. **49**(6), 2343–2353 (2011)
82. Shabou, A., Tupin, F.: A Markovian approach for DEM estimation from multiple InSAR data with atmospheric contributions. IEEE Geosci. Remote Sens. Lett. **9**(4), 764–768 (2012)
83. Shabou, A., Baselice, F., Ferraioli, G.: Urban digital elevation model reconstruction using very high resolution multichannel insar data. IEEE Trans. Geosci. Remote Sens. **50**(11), 4748–4758 (2012)

84. Aharon, M., Elad, M., Bruckstein, A.: K-SVD: an algorithm for designing overcomplete dictionaries for sparse representation. IEEE Trans. Signal Process. **54**(11), 4311–4322 (2006)
85. Pati, Y.C., Rezaiifar, R., Krishnaprasad, P.: Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition. In: Conference Record of The Twenty-Seventh Asilomar Conference on Signals, Systems and Computers, pp. 40–44. IEEE (1993)
86. Elad, M., Aharon, M.: Image denoising via sparse and redundant representations over learned dictionaries. IEEE Trans. Image Process. **15**(12), 3736–3745 (2006)
87. Roth, S., Black, M.J.: Fields of experts: a framework for learning image priors. In: IEEE Conference on Computer Vision and Pattern Recognition **2**, 860–867 (2005)
88. Roth, S., Black, M.J.: Fields of experts. Int. J. Comput. Vis. **82**(2), 205–229 (2009)
89. Elad, M., Milanfar, P., Rubinstein, R.: Analysis versus synthesis in signal priors. Inverse problems **23**(3), 947 (2007)
90. Chen, Y., Feng, W., Ranftl, R., Qiao, H., Pock, T.: A higher-order MRF based variational model for multiplicative noise reduction. IEEE Signal Process. Lett. **21**(11), 1370–1374 (2014)
91. Yu, G., Sapiro, G., Mallat, S.: Solving inverse problems with piecewise linear estimators: from Gaussian mixture models to structured sparsity. IEEE Trans. Image Process. **21**(5), 2481–2499 (2012)
92. Zoran, D., Weiss, Y.: From learning models of natural image patches to whole image restoration. In: IEEE International Conference on Computer Vision, pp. 479–486 (2011)
93. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. Ser. B (Methodol.), pp. 1–38 (1977)
94. Tabti, S., Deledalle, C.A., Denis, L., Tupin, F.: Modeling the distribution of patches with shift-invariance: application to SAR image restoration. In: IEEE International Conference on Image Processing, pp. 96–100 (2014)
95. Xie, H., Pierce, L.E., Ulaby, F.T.: SAR speckle reduction using wavelet denoising and Markov random field modeling. IEEE Trans. Geosci. Remote Sens. **40**(10), 2196–2212 (2002)
96. Foucher, S.: SAR image filtering via learned dictionaries and sparse representations. In: IEEE International Geoscience and Remote Sensing Symposium, vol. 1, pp. I–229 (2008)
97. Mäkitalo, M., Foi, A., Fevralev, D., Lukin, V.: Denoising of single-look SAR images based on variance stabilization and nonlocal filters. In: IEEE International Conference on Mathematical Methods in Electromagnetic Theory, pp. 1–4 (2010)
98. Lee, J.S.: Digital image smoothing and the sigma filter. Comput. Vis. Graphics Image Process. **24**(2), 255–269 (1983)
99. Buades, A., Coll, B., Morel, J.M.: A review of image denoising algorithms, with a new one. Multiscale Model. Simul. **4**(2), 490–530 (2005)
100. Conradsen, K., Nielsen, A.A., Schou, J., Skriver, H.: A test statistic in the complex Wishart distribution and its application to change detection in polarimetric SAR data. IEEE Trans. Geosci. Remote Sens. **41**(1), 4–19 (2003)
101. Deledalle, C.A., Denis, L., Tupin, F.: How to compare noisy patches? Patch similarity beyond Gaussian noise. Int. J. Comput. Vis. **99**(1), 86–102 (2012)
102. Deledalle, C.A., Denis, L., Tupin, F., Reigber, A., Jäger, M.: NL-SAR: a unified nonlocal framework for resolution-preserving (Pol)(In) SAR Denoising. IEEE Trans. Geosci. Remote Sens. **53**(4), 2021–2038 (2015)
103. Polzehl, J., Spokoiny, V.: Propagation-separation approach for local likelihood estimation. Probab. Theory Relat. Fields **135**(3), 335–362 (2006)
104. Kay, S.M.: Fundamentals of Statistical Signal Processing, vol. I: Estimation Theory (1993)
105. Lee, J.S., Wen, J.H., Ainsworth, T.L., Chen, K.S., Chen, A.J.: Improved sigma filter for speckle filtering of SAR imagery. IEEE Trans. Geosci. Remote Sens. **47**(1), 202–213 (2009)
106. Dabov, K., Foi, A., Katkovnik, V., Egiazarian, K.: Image denoising by sparse 3-D transform-domain collaborative filtering. IEEE Trans. Image Process. **16**(8), 2080–2095 (2007)
107. Lebrun, M., Buades, A., Morel, J.M.: A nonlocal Bayesian image denoising algorithm. SIAM J. Imaging Sci. **6**(3), 1665–1688 (2013)

108. Parrilli, S., Poderico, M., Angelino, C.V., Verdoliva, L.: A nonlocal SAR image denoising algorithm based on LLMMSE wavelet shrinkage. IEEE Trans. Geosci. Remote Sens. **50**(2), 606–616 (2012)
109. Deledalle, C.A., Denis, L., Ferraioli, G., Tupin, F.: Combining patch-based estimation and total variation regularization for 3D InSAR reconstruction. In: IEEE International Geoscience and Remote sensing Symposium, Milan, Italy (2015)
110. Lee, J.S., Pottier, E.: Polarimetric Radar Imaging: Basics and Applications. CRC Press (2009)
111. Massonnet, D., Souyris, J.C.: Imaging with Synthetic Aperture Radar. EPFL Press (2008)
112. Argenti, F., Lapini, A., Bianchi, T., Alparone, L.: A tutorial on speckle reduction in synthetic aperture radar images. IEEE Trans. Geosci. Remote Sens. **1**(3), 6–35 (2013)
113. Deledalle, C.A., Denis, L., Poggi, G., Tupin, F., Verdoliva, L.: Exploiting patch similarity for SAR image processing: the nonlocal paradigm. IEEE Signal Process. Mag. **31**(4), 69–78 (2014)

Chapter 5
Polarimetric SAR Modelling: Mellin Kind Statistics and Time-Frequency Analysis

Torbjørn Eltoft, Laurent Ferro-Famil, Stian N. Anfinsen and Anthony P. Doulgeris

Abstract Polarimetric synthetic aperture radar (PolSAR) remote sensing refers to measurement techniques that exploit the ability of a back scattering target to transform the polarization state of incoming electromagnetic waves to extract information. This field of remote sensing deals with complex, multi-dimensional data sets, requiring multi-dimensional signal analysis strategies. Target information is in general extracted from PolSAR data by an analysis approach referred to as polarimetric target decompositions, a field of research which has been successfully developed over the last few decades. Low resolution SAR images of distributed natural scenes contain a high number of scatterers within each resolution cell, resulting in Gaussian signal statistics. Gaussian signal statistics and target decompositions found the basis for classical analysis of PolSAR data. With the improving spatial resolution of currently operating SARs, the Gaussian assumption is frequently challenged and often abandoned, in particular for scenes of urban environment, but also for natural surfaces such as forest and sea. This development has stimulated research on non-Gaussian models for representing the statistics of SAR and PolSAR signals. In particular, non-Gaussian modelling of multi-dimensional SAR signals using the doubly stochastic product model has received an increased attention in recent years. The chapter will give a brief introduction to techniques for analysis of PolSAR data using statistical signal processing. Here two axes will be explored: one that focuses on non-Gaussian signals statistics, including sections on modelling, estimation, and classification; and one that focuses on the signals time-frequency properties, discussing target detection and discrimination.

T. Eltoft (✉) · S.N. Anfinsen · A.P. Doulgeris
UiT The Arctic University of Norway, Tromsø, Norway
e-mail: Torbjorn.Eltoft@uit.no

S.N. Anfinsen
e-mail: Stian.Normann.Anfinsen@uit.no

A.P. Doulgeris
e-mail: Anthony.P.Doulgeris@uit.no

L. Ferro-Famil
University of Rennes 1, Rennes, France
e-mail: Laurent.Ferro-Famil@univ-rennes1.fr

G. Moser and J. Zerubia (eds.), *Mathematical Models for Remote Sensing Image Processing*, Signals and Communication Technology,
https://doi.org/10.1007/978-3-319-66330-2_5

5.1 Introduction

A Polarimetric Synthetic Aperture Radar (PolSAR) is a coherent imaging system that achieves a high spatial resolution by coherently combining many individual pulsed signals into one effective aperture as the sensor moves over the target area. These are systems where images are produced by "object illumination with the particular property that the phasor amplitudes at all object points vary in unison. Thus, while any two objects may have different fixed relative phases, their absolute phases are varying in time in identical fashions" [1]. A common feature of coherent imagery is the presence of a signal phenomenon known as *speckle*. Speckle is generally explained to be attributed to the random interference of many coherent wave components reflected from different microscopic elements of the rough surface. Rough will in this regard refer to coarseness of the surface layer that is large compared to the wave length of the illumination. At a distant observation point, the various de-phased coherent components will add and give constructive or destructive interference, which will give the image a noisy, granular appearance. Modelling the signal statistics of coherent images has been a research area for several decades. There are several reasons why this research area is important: Speckle is an consequence of the image formation process in coherent imagery; it masks details and may complicate interpretations. Hence, de-speckling, itself, is a goal, and some of the existing speckle filter algorithms lend themselves on a Bayesian approach in which case accurate local statistical models for the speckled and speckle-free images are essential [2, 3]. On the other hand, speckle is an intrinsic property of the signal, and as such it carries important information of the scattering target medium. The probability model contains this information.

The first part of this chapter is devoted to statistical modelling of focussed SAR and PolSAR images. We give a brief review of classical models for signal parameters like the amplitude and in the intensity under the Gaussian signal assumption, and discuss how these models can easily be adapted to a non-Gaussian regime under the product model. We start by describing parametric models associated to single channel SAR parameters, and subsequently extend the description to the multivariate polarimetric case. This part of the chapter also includes a section on parameter estimation and Mellin kind statistics (also referred to as second kind statistics, and briefly mentioned in Chap. 4), which has turned out to be particular efficient and useful when the signal models are formulated as compound models. The final section of part one, demonstrates how the statistical models can be used to enhance information retrieval from PolSAR data. The demonstration includes examples in supervised classification and unsupervised segmentation of a full-polarimetric SAR scene.

The second part of the chapter describes a completely different approach to analysis of polSAR data and delves deeper into the multi-pulsed construction of the synthetic aperture signal. A time-frequency characterisation of complex polarimetric features uncovers potential temporal variations of the signal during the multi-pulsed SAR acquisition that affect feature estimates derived from the SAR/PolSAR data. This section presents different techniques for detecting scatterers having a varying

response during the SAR acquisition, and demonstrates how the underlying physical phenomenon can be characterised based on specific coherent Time-Frequency (TF) decompositions applied on focused images. Both sections involve strong mathematical and statistical modelling that can be used separately or combined to better interpret SAR and PolSAR images. Note that some of the basic material on SAR imaging and speckle modelling in this chapter have some overlap with Chap. 5. We have decided to include this for completeness.

5.2 Signal Modelling

Goodman [2] proposed the first statistical model for single-look, single polarization SAR data in which the measured signal at each pixel in a SAR image is the vector sum of backscatter from a multitude of individual scatterers. Hence, we may write the detected signal from a given resolution cell as

$$S = X + jY = A\exp(j\Phi) = \sum_{k=1}^{N} a_k \exp(j\phi_k), \tag{5.1}$$

where N is the number of scatterers in the cell, a_k is the amplitude and ϕ_k is the phase of the k-th scatterer. X and Y are referred to as the in-phase (I) and quadrature (Q) components, respectively. The scattering medium is modelled as a distribution of mutually independent, statistically identical elementary scatterers.

In a locally homogeneous area (or volume), the population of random scatterers can be modelled as a compound Poisson process, in which the scatterers are uniformly distributed in a large measurement space. The probability that a resolution cell contains N scatterers obeys a Poisson probability law with mean number $\bar{N}$, i.e.,

$$p(N) = \frac{\bar{N}^N}{N!}\exp(-\bar{N}). \tag{5.2}$$

As the size of the resolution cells increases, implying that $\bar{N} \to \infty$, the probability of S in (1) becomes a complex Gaussian distribution. This is a consequence of the central limit theorem, and is known as *fully developed speckle*. Speckle is considered to be fully developed under the following conditions:

- The imaged surface is rough compared to the wavelength of the incident electromagnetic energy,
- A large number of independent scattering elements contribute to the measured signal in a single resolution cell,
- There are no dominant scatterers in any resolution cell.

Under the assumption of fully developed speckle, we can further deduce that the complex return from each resolution cell is a circularly symmetric Gaussian random variable, meaning that X and Y in (5.1) are statistically independent and identically distributed with zero mean. The corresponding distribution for the magnitude, i.e., $A = \sqrt{X^2 + Y^2}$, is modelled by the Rayleigh probability density function

$$p(A;\sigma) = \frac{A}{\sigma^2} \exp\left(-\frac{A^2}{2\sigma^2}\right), \tag{5.3}$$

with σ^2 being the variance of X and Y.[1] It readily also follows that the intensity or power, $I = X^2 + Y^2$, obeys the negative exponential distribution (e.g., [3]):

$$p(I;\sigma) = \frac{I}{2\sigma^2} \exp\left(-\frac{I}{2\sigma^2}\right). \tag{5.4}$$

We note that the average intensity is $\mathscr{E}\{I\} = 2\sigma^2$. In a distributed target, the observed power is an estimate of an underlying radar cross section (RCS), whose actual value is hidden by the interference between the individual scattering contributions [4]. In high-resolution radars, the backscattered signal will deviate from the Gaussian models. Non-Gaussian distributions are commonly observed, such as when the number of scatterers in a resolution cell is small, the scatterers are organised with some kind of periodicity, or when there are some dominant target components present in the cell. Also, when the signals can be considered to be a mixture of contributions from several distributed targets, the resulting statistical signal model will be non-Gaussian.

5.3 The Product Model

There are a variety of distributions that attempts to model non-Gaussian signal statistics. One of the most commonly used non-Gaussian models is known as the K-model. Its associated magnitude distribution was first introduced to describe the spatial distribution of certain larvae in terms of a two-dimensional (2-D) random walk model, coupled with an exponential stopping time distribution [5]. This result was generalized by Yasuda [6] by using a Γ-distributed stopping time in a Rayleigh random walk of Brownian motion type. The K model has subsequently been used as a generic model for scattered radiation [7, 8], and non-Rayleigh microwave speckle

[1] Note that in this Chapter, we use the notation $p(X;\theta)$ to denote the pdf of a random variable X, parametrised by θ, and $p(X|\tau)$ to denote the pdf of X conditioned on τ. Both upper and lower case letters are used to denote a random variable.

[9–11]. The K-model is an example of a non-Gaussian distribution that can be generated according to the so-called product model. The product model represents a simple, generic method for generating non-Gaussian distributions with heavy tails. It is theoretically founded on a paper by Andrews and Mallow [12], where it was shown that if the probability density function (pdf) of some random variable Y, $p(y)$, is symmetric about zero, and the derivatives of $p(y)$ satisfy

$$\left(-\frac{d}{dy}\right)^k p(y) \geq 0 \quad \text{for } y > 0, \tag{5.5}$$

then there exist independent variables X and τ, with X being a standard normal variable, such that

$$Y = \sqrt{\tau} X. \tag{5.6}$$

The variable τ is allowed to take on only positive values. A random variable Y, which can be expressed as in (5.6), accordingly is a product of two independent variables, hence the name *product model*. It is also referred to as *a normal variance mixture model*, or *a scale mixture of Gaussians*. If the mean of Y is non-zero, (5.6) may be modified by adding a scalar μ corresponding to the actual mean value. The marginal pdf of Y is accordingly obtained by integrating the conditional distribution $p(y|\tau)$ over $p(\tau; \boldsymbol{\theta})$, i.e.,

$$p(y) = \int_0^\infty p(x|\tau) p(\tau; \boldsymbol{\theta}) d\tau, \tag{5.7}$$

where $\boldsymbol{\theta}$ is a parameter vector associated with the pdf of the texture. As mentioned in Chap. 5, the product model in Eq. (5.6) can easily be extended to the complex radar signal in (5.1), in which case we write

$$S = \sqrt{\tau} X + j\sqrt{\tau} Y. \tag{5.8}$$

The stochastic variable τ is here referred to as the *radar texture*, and is used to describe the case when the radar cross section of the target surface is inhomogeneous. This inhomogeneity will affect the signal statistics in such a way that the resulting distribution becomes non-Gaussian, with a heavy tail. It is readily seen that the quadrature components of S still maintain symmetry in the sense that they are uncorrelated with the same pdf, and conditioned on τ, S is still a circular symmetric Gaussian random variable. It is also easy to verify that the magnitude and intensity distributions are given as:

$$p(A; \sigma, \boldsymbol{\theta}) = \int_0^\infty \frac{A}{\tau\sigma^2} \exp\left(-\frac{A^2}{2\tau\sigma^2}\right) p(\tau; \boldsymbol{\theta}) d\tau, \tag{5.9}$$

and

$$p(I;\sigma,\boldsymbol{\theta}) = \int_0^\infty \frac{I}{2\tau\sigma^2}\exp\left(-\frac{I}{2\tau\sigma^2}\right)p(\tau;\boldsymbol{\theta})d\tau, \tag{5.10}$$

respectively.

In the remainder, we will assume that the expected value $\mathscr{E}\{\tau\} = 1$, such that the average power in the radar signal is not altered by the introduction of the texture term. We will also use $R = \mathscr{E}\{I\} = 2\sigma^2$ to denote the *average signal intensity*.

In *multilooking*, which is the simplest way to de-speckle a SAR image, the intensity is averaged over several pixels within a window centred on a specific pixel. When averaging L pixels, we will assume that all signals result from the same underlying RCS, and that the pixel intensities are independent variables. We then get a Γ-distributed multilooked intensity with a pdf given as

$$p(I;L,R) = \frac{L^L I}{R^L\Gamma(L)}\exp\left(-\frac{LI}{R}\right). \tag{5.11}$$

When the product model is applied to describe multilooked data, the assumption is that the RCS is constant within the averaging window, and hence the textured multilook intensity distribution is obtained by using Eq. (5.11) in the integration over the texture domain. In this case, for a given τ, the local average intensity will be τR. This integral is in many cases possible to express in closed form. Table 5.1 lists the pdfs for Γ and Fischer-distributed texture.

Table 5.1 Analytic expressions for the integrals defining the complete pdfs corresponding to Gamma, and Fisher texture models. The letters $\mathscr{B}(x)$, $K_\alpha(x)$, and $U(a,b,x)$, refer to the Beta function, Modified Bessel function of second kind and order α, and Kummer's U function with parameters a and b, respectively

Model	Texture pdf: $f(\tau;\boldsymbol{\theta})$	Compound pdf: $P(I;L,\boldsymbol{\theta}) = \int_0^\infty P(I\|\tau)f(\tau;\boldsymbol{\theta})dt$
No texture	$p_T(\tau) = \delta(\tau-1)$	$p(I;L,R) = \frac{L^L I}{R^L\Gamma(L)}\exp(-\frac{LI}{R})$.
Gamma	$p_T(\tau;\alpha) = \frac{\alpha^\alpha}{\Gamma(\alpha)}t^{\alpha-1}\exp(-\alpha t)$	$p(I;L,R,\alpha) = \frac{2}{\Gamma(\alpha)\Gamma(L)}\left(\frac{L\alpha}{R}\right)^{\frac{L+\alpha}{2}} I^{\frac{L+\alpha}{2}-1} \times K_{\alpha-L}\left(2\sqrt{\frac{\alpha LI}{R}}\right)$
Fisher	$f_T(\tau;\alpha,\lambda) = \mathscr{B}^{-1}(\alpha,\lambda)\frac{\alpha}{\lambda-1}\frac{\left(\frac{\alpha\tau}{\lambda-1}\right)^{\alpha-1}}{\left(\frac{\alpha\tau}{\lambda-1}+1\right)^{\alpha+\lambda}}$	$p(I;L,R,\alpha,\lambda) = \mathscr{B}^{-1}(\alpha,\lambda)\left(\frac{L}{R}\right)^L\frac{\Gamma(L+\lambda)}{\Gamma(L)}\left(\frac{\alpha}{\lambda-1}\right)^L \times I^{(L-1)}U(L+\lambda, L-\alpha+1, \frac{\alpha}{\lambda-1}\frac{LI}{R})$

5.4 Radar Polarimetry

Radar polarimetry deals with the full vector nature of electromagnetic waves. General introductory texts on radar and SAR polarimetry are found in e.g., [13–15]. When the electromagnetic wave passes through a medium of changing index of refraction, or when it interact with an object or a target surface and is reflected or scattered, the characteristic information about the reflectivity, shape and orientation of the reflecting body can be obtained by polarimetric analysis of the echoes [15]. This information is only available if the radar system has full polarimetric capability. For the linear polarization basis, this means the system is able to measure the backscattered signal in four polarization channels. For example, in the horizontal and vertical polarization basis, the four combinations of channels are HH, HV, VH, VV. This is mathematically formulated by means of the Sinclair matrix (also referred to as the scattering matrix), which relates *the Jones vector* of the backscattered wave to the Jones vector of the incident wave, as shown in (5.12) below

$$\begin{bmatrix} E_h^s \\ E_v^s \end{bmatrix} = \frac{\exp(jkr)}{r} \begin{bmatrix} S_{HH} & S_{HV} \\ S_{VH} & S_{VV} \end{bmatrix} \begin{bmatrix} E_h^i \\ E_v^i, \end{bmatrix}. \tag{5.12}$$

Here k is wavenumber of the incident electromagnetic wave, r is the distance between the radar antenna and target, E_j^i and E_j^s, $j \in \{h, v\}$, denote the jth linear polarization component of the incident and scattered waves, respectively, and the entries of the S-matrix, i.e., S_{HH}, S_{HV}, S_{VH}, S_{VV}, define the channel wise complex scattering coefficients. For practical analysis of polarimetric data, the S-matrix is often vectorized; either as a lexicographic *scattering or target vector*

$$\mathbf{s} = [S_{HH}, S_{HV}, S_{VH}, S_{VV}]^t, \tag{5.13}$$

or as a *Pauli scattering vector*

$$\mathbf{k} = \frac{1}{\sqrt{2}}[S_{HH} + S_{VV}, S_{HH} - S_{VV}, S_{HV} + S_{VH}, j\,(S_{HV} - S_{VH})]^t, \tag{5.14}$$

where the superscript t denotes transpose, and the square-root factor in (5.14) maintains the total backscattered power. Data in this format is denoted single look complex (SLC) polarimetric data. A scattering mechanism is defined as a normalized Pauli scattering vector, and is used to characterize differences in polarized wave scattering [14].

Statistically, the polarimetric target vector, following the description in the previous section, would be represented as a multivariate complex statistical variable. Under the assumption of fully developed speckle, $\mathbf{s}$ (or $\mathbf{k}$) would be multivariate complex Gaussian, and the distribution given as:

$$p(\mathbf{s}) = \frac{1}{\pi|\Sigma|} \exp(-\mathbf{s}^\dagger \Sigma^{-1} \mathbf{s}), \tag{5.15}$$

where Σ defines the complex covariance matrix of $\mathbf{s}$, i.e., $\Sigma = \mathscr{E}\{\mathbf{s}\mathbf{s}^{\dagger}\}$, $|\cdot|$ is the matrix determinant, and the superscript $\dagger$ refers to conjugate transpose. The individual elements of the polarimetric covariance matrix carries information about channel-wise correlations, and is an important source of information. If the speckle is not fully developed, we will have deviation from Gaussian statistics, and we may adapt the vector form of the product model to describe the statistics. Here, we will only consider the simplest version of the multivariate product model, where it is assumed that the radar texture is equal in all polarisation channels. The model for $\mathbf{s}$ then takes the form:

$$\mathbf{s} = \sqrt{\tau}\ [S_{HH}, S_{HV}, S_{VH}, S_{VV}]^t, \tag{5.16}$$

where τ would be a scalar stochastic variable, with the same properties as above. Multilooking in the polarimetric case is implemented using the outer product of the $\mathbf{s}$ or $\mathbf{k}$ vectors, and results in the sample *covariance matrix* or sample *coherence matrix*. These matrices are hence defined as

$$\mathbf{C} = \frac{1}{L}\sum_{i=1}^{L}\mathbf{s}_i\mathbf{s}_i^{\dagger} \ \text{ and } \ \mathbf{T} = \frac{1}{L}\sum_{i=1}^{L}\mathbf{k}_i\mathbf{k}_i^{\dagger}, \tag{5.17}$$

respectively, where L is the nominal number of looks being averaged. This data format is known as multilook complex polarimetric data. When multilooking L pixels, we will assume that all signals result from the same underlying channel-wise RCSs, i.e., the pixel-wise scattering matrices are identical distributed, independent random variables.

In the fully developed speckle case, in which case $\mathbf{s}$ is a multivariate Gaussian variable, the sample covariance matrix $\mathbf{C}$ will be *scaled complex Wishart* distributed, with a pdf given as

$$p(\mathbf{C};\, L;\, \Sigma) = \frac{L^{Ld}|\mathbf{C}|^{L-d}}{I(L,d)|\Sigma|^{L}}\exp(-L\,\mathrm{tr}(\Sigma^{-1}\mathbf{C})). \tag{5.18}$$

In (5.18), Σ is the true covariance of $\mathbf{s}$, d is the dimension of $\mathbf{s}$, which in general is 4, (but is 3 if reciprocity can be assumed), $\mathrm{tr}(\cdot)$ is the trace of a matrix, and $I(L,d) = \pi^{\frac{d(d-1)}{2}}\prod_{i=0}^{d-1}\Gamma(L-i)$ is a normalisation constant.

Again, when the product model is applied to describe multilook polarimetric data, the assumption is that the RCSs are constant within the averaging window, and hence the textured multilook covariance matrix is given as

$$\mathbf{C} = \tau\mathbf{W}, \tag{5.19}$$

where τ is the texture, and $\mathbf{W}$ here refers to complex scaled Wishart distributed speckle. The pdf of $\mathbf{C}$ is then obtained by integrating over the texture distribution.

Table 5.2 Analytic expressions for the integrated product models of (5.19) corresponding to Γ and Fischer-distributed texture. As for Table 5.1, the letters $\mathscr{B}(x)$, $K_\alpha(x)$, and $U(a,b,x)$, refer to the Beta function, Modified Bessel function of second kind and order α, and Kummer's U function with parameters a and b, respectively

Model	Texture pdf, $f(\tau)$	Compound pdf: $p(\mathbf{C};L;\Sigma,\boldsymbol{\theta}) = \int_0^\infty p(\mathbf{C};L,\Sigma\|\tau)f(\tau;\boldsymbol{\theta})d\tau$
No texture	$f(\tau) = \delta(\tau-1)$	$p(\mathbf{C};L;\Sigma) = \frac{L^{Ld}\|\mathbf{C}\|^{L-d}}{I(L,d)\|\Sigma\|^L}\exp(-L\,\mathrm{tr}(\Sigma^{-1}\mathbf{C}))$
Gamma	$f(\tau;\alpha) = \frac{\alpha^\alpha}{\Gamma(\alpha)}\tau^{\alpha-1}\exp(-\alpha t)$	$p(\mathbf{C};L,\Sigma,\alpha) = \frac{2\,\|\mathbf{C}\|^{L-d}(L\alpha)^{\frac{\alpha+Ld}{2}}}{\|\Sigma\|^L I(L,d)\Gamma(\alpha)}\left(\mathrm{tr}(\Sigma^{-1}\mathbf{C})\right)^{\frac{\alpha-Ld}{2}} \times K_{\alpha-Ld}(2\sqrt{L\alpha\,\mathrm{tr}(\Sigma^{-1}\mathbf{C})})$
Fisher	$f_T(t;\alpha,\lambda) = \mathscr{B}^{-1}(\alpha,\lambda)\frac{\alpha}{\lambda-1}\frac{\left(\frac{\alpha\tau}{\lambda-1}\right)^{\alpha-1}}{\left(\frac{\alpha\tau}{\lambda-1}+1\right)^{\alpha+\lambda}}$	$p(\mathbf{C};L,\Sigma,\alpha,\lambda) = \mathscr{B}^{-1}(\alpha,\lambda)\frac{L^{Ld}\|\mathbf{C}\|^{L-d}}{I(L,d)\|\Sigma\|^L}\left(\frac{\alpha}{(\lambda-1)}\right) \times\Gamma(Ld+\lambda)\,U(Ld+\lambda, Ld-\alpha+1, \frac{L\,\mathrm{tr}(\Sigma^{-1}\mathbf{C})\alpha}{\lambda-1})$

These integrals are in many cases possible to express in closed form. Table 5.2 lists the pdfs for Γ- and Fischer-distributed texture.

5.5 Parameter Estimation

Section 5.3 presents the product model for single-polarisation amplitude and intensity and Sect. 5.4 extends it to a product model for the polarimetric sample covariance matrix. Depending on the choice of distribution for the texture parameter, this leads to different distributions for the compound observable. In order to utilise these distributions in model-based image analysis, such as classification, segmentation, change detection or target detection, the distribution parameters must be estimated. This is the scope of this section. It is assumed that the reader is familiar with maximum likelihood estimation and the method-of-moments estimation.

Recall that the product model decomposes into two terms. The first term represents fully developed speckle, or stochastic variations due to the interference phenomenon which occurs when a theoretically infinite number of coherent radar echoes are summed. The second term represents additional signal variation, which could stem from a number of sources, such as a finite and fluctuating number of scatterers or local variations in the radar reflectivity. These effects can be caused by various circumstances relating to sensor configurations or the imaged environment. Regardless of the origin, the second term is referred to as texture.

In the multilook polarimetric product model, $\mathbf{C} = \tau \cdot \mathbf{W}$, the scaled complex Wishart distributed speckle matrix, $\mathbf{W}$, is parametrised by two parameters: the equivalent number of looks, L, and the mean covariance matrix, $\boldsymbol{\Sigma}$. The texture variable,

τ, is generally parametrised by a parameter vector, $\boldsymbol{\theta}$. We shall look at some concrete examples of texture parameters, after first discussing estimation of $\boldsymbol{\Sigma}$ and L.

5.5.1 Covariance Estimation

The population covariance $\boldsymbol{\Sigma}$ can either be estimated from a collection of scattering vectors or a collection of sample covariance or sample coherency matrices, depending on which format is available. The vector sample is denoted $\mathscr{S} = \{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$ when provided in the lexicographic basis and $\mathscr{K} = \{\mathbf{k}_1, \ldots, \mathbf{k}_n\}$ in the Pauli basis. The corresponding matrix samples are $\mathscr{C} = \{\mathbf{C}_1, \ldots, \mathbf{C}_n\}$ and $\mathscr{T} = \{\mathbf{T}_1, \ldots, \mathbf{T}_n\}$, respectively.

5.5.1.1 Sample Mean Estimator for Matrix Sample

We have assumed that the texture variable is normalised, in the sense that $E\{\tau\} = 1$. By use of the independence of τ and $\mathbf{W}$, it follows that $E\{\mathbf{C}\} = \boldsymbol{\Sigma}$, as a general result irrespective of the distribution of τ. This proves that $\boldsymbol{\Sigma}$ can be estimated for all distributions of τ with the well-known sample mean estimator (SME). Given a sample $\mathscr{C}$, the SME for the matrix sample $\mathscr{C}$ is defined as

$$\hat{\boldsymbol{\Sigma}}_{SME} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{C}_i \,. \tag{5.20}$$

The SME for the matrix sample $\mathscr{T}$ is identical, since the coherency matrix $\mathbf{T}$ is simply the sample covariance of the scattering vector $\mathbf{k}$ on Pauli basis format.

5.5.1.2 Sample Mean Estimator for Vector Sample

The SME for a vector sample $\mathscr{S}$ is similarly given as

$$\hat{\boldsymbol{\Sigma}}_{SME} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{s}_i \mathbf{s}_i^{\dagger} \,, \tag{5.21}$$

and equivalently for the sample $\mathscr{K}$ in the Pauli basis, with $\mathbf{k}_i$ replacing $\mathbf{s}_i$.

In the case of a circular complex Gaussian vector sample, meaning that $\tau \equiv 1$ and $p_\tau(t) = \delta(t-1)$, then the SME is known to be the maximum likelihood estimator (MLE) of $\boldsymbol{\Sigma}$. In the general case, the formulation of a MLE is complicated and it does not have an analytic expression. Gini and Greco [16] derived the general MLE for the population covariance based on the product model for the complex scattering vector. It is given as

$$\hat{\boldsymbol{\Sigma}}_{MLE}^{(k+1)} = \frac{1}{n}\sum_{i=1}^{n} \frac{h_{d+1}\left(\mathbf{s}_i^{\dagger}(\hat{\boldsymbol{\Sigma}}_{MLE}^{(k)})^{-1}\mathbf{s}_i\right)}{h_d\left(\mathbf{s}_i^{\dagger}(\hat{\boldsymbol{\Sigma}}_{MLE}^{(k)})^{-1}\mathbf{s}_i\right)} \cdot \mathbf{s}_i\mathbf{s}_i^{\dagger}\,, \tag{5.22}$$

where $k \in \{1, \ldots, K\}$ is the iteration number and K is the number of iterations, d is the dimension of the scattering vector, and $h_d(q)$ is the so-called nonlinear memoryless function, which is defined as

$$h_d(q) = \int_0^{+\infty} t^{-d} \exp\left(-\frac{q}{t}\right) p_\tau(t)\, dt\,, \tag{5.23}$$

and which involves the texture distribution $p_\tau(t)$.

Equation (5.22) is an iterative solution of a transcendental equation. The estimate at the $(k+1)$th step is obtained from the estimate at step k. The process requires prior knowledge of $p_\tau(t)$, which includes the texture parameter vector $\boldsymbol{\theta}$, showing that this becomes in practice a compound estimation problem where $\boldsymbol{\Sigma}$ and $\boldsymbol{\theta}$ must be estimated simultaneously. The iterative estimator must also be initialised with a first guess, which can be provided by the SME.

As an example, let τ be a unit-mean Gamma distributed random variable. Then the pdf of $\mathbf{C}$ becomes the matrix-variate K distribution given in Table 5.2. After evaluating Eq. (5.23), it is found [16] that the MLE for $\boldsymbol{\Sigma}$ becomes

$$\hat{\boldsymbol{\Sigma}}_{MLE}^{(k+1)} = \frac{1}{n}\sum_{i=1}^{n} \sqrt{\frac{\alpha}{q_i}} \frac{K_{\alpha-d-1}(\sqrt{4\alpha q_i})}{K_{\alpha-d}(\sqrt{4\alpha q_i})} \cdot \mathbf{s}_i\mathbf{s}_i^{\dagger}\,, \tag{5.24}$$

where

$$q_i = \mathbf{s}_i^{\dagger}(\hat{\boldsymbol{\Sigma}}_{MLE}^{(k)})\mathbf{s}_i\,. \tag{5.25}$$

With Gamma distributed texture, the MLE is seen to contain the special function $K_\nu(q)$, known as the modified Bessel function of the second kind with order ν. This gives an estimator with high computational cost. In addition, we note that the shape parameter α associated with the texture variable is assumed known, which it is not in most practical cases. The MLE is therefore seldom used in practice.

5.5.1.3 Robust M-Estimator

Due to the practical constraints of the MLE, the SME is often preferred also with data that are known to contain texture. For moderate to low texture, meaning that the shape parameter of the texture variable is relatively high, the performance loss of the SME is small, as shown by Tao et al. in [17]. If the texture is significant, then the so-called M-estimator can be an alternative.

The M-estimator belongs to Huber's general class of robust maximum likelihood type estimators, hence the name. This approach was first applied to covariance estimation by Tyler [18], and has later become popular in radar image analysis, after it was rediscovered in [16, 19]. The M-estimator for $\boldsymbol{\Sigma}$ is defined as

$$\hat{\boldsymbol{\Sigma}}_M^{(k+1)} = \frac{d}{n} \sum_{i=1}^{n} \frac{\mathbf{s}_i \mathbf{s}_i^\dagger}{\mathbf{s}_i^\dagger (\boldsymbol{\Sigma}_M^{(k)})^{-1} \mathbf{s}_i} . \tag{5.26}$$

As seen, the M-estimator must also be computed iteratively. However, it does not require any prior information about the texture distribution. The computation is also mathematically simple, which makes it a practical alternative to the SME. For these reasons, a large interest in this estimator has been observed in the radar literature, where it is commonly referred to as the fixed-point estimator.

Pascal et al. proved the existence and uniqueness of the M-estimator solution, the convergence of its recursive scheme under any initialisation [20], and that it is asymptotically complex Wishart distributed [21]. Many applications to analysis of PolSAR data have followed, such as [22]. Despite the popularity, it has been demonstrated [17] that the M-estimator suffers a significant performance loss in low to moderate texture and only supersedes the SME when the texture level is higher than what may be realistic to find in real PolSAR data. The robustness of the M-estimator to outliers was not considered in this study.

5.5.2 Mellin Kind Statistics

As shown in previous sections, the doubly stochastic product model leads to mathematically complicated distributions. Hence, it becomes difficult to use a maximum likelihood approach to estimate the population covariance matrix. The same is true when we try to estimate the shape parameters of the texture distributions, referred to as *texture parameters* of the overall distribution for $\mathbf{C}$. The common solution has been to resort to moment-based estimators.

The product model decomposes the polarimetric covariance (or coherency) matrix into two factors, speckle and texture, that are statistically independent. It would be desirable if we could find moment expressions that also separate the contribution of speckle and texture, so that we can easily isolate speckle parameters from texture parameters. As mentioned in Chap. 4, it just so happens that this is possible. It requires that we use logarithmic moments. It also proves helpful to introduce the Mellin transform to derive the theoretical foundation for logarithmic statistics.

To provide some intuition about the appropriateness of logarithmic statistics in this context, recall that the logarithm transforms the product model into an additive one. From basic signal processing and linear system theory, we have the familiar additive signal model with contributions from a statistically independent signal and noise component. For this model, a certain type of moment expressions are known to possess the desired property that moments of signal and noise decompose additively,

namely cumulants. Cumulants can be retrieved from the so-called cumulant generating function, which is related to the pdf by the Laplace transform or, equivalently, by the Fourier transform.

Moreover, note that the logarithmic sample mean is a sufficient statistic for the shape parameter of a Gamma distributed population, meaning that it contains the maximum amount of information about the shape parameter which can be retrieved from any data sample drawn from this distribution. Even though the product model modulates the Gamma distributed intensity or complex Wishart distributed sample covariance with a texture variable, it still makes sense that logarithmic statistics of the data should hold much information about the shape parameters involved. We shall in the following see that the Mellin transform is effectively a Laplace transform computed on logarithmic scale, and that it can be used to define logarithmic cumulants.

5.5.2.1 The Mellin Transform

Before we indulge in logarithmic statistics, we must define the Mellin transform, both for the univariate and the matrix-variate case. The Mellin transform is lesser known than its close relatives, the Fourier and the Laplace transforms. Nonetheless, its use of a polynomial transform kernel instead of an exponential one provides it with a scale invariance property which has found many important applications in signal processing, physics and other fields.

The Mellin transform of the real-valued function $f(x)$ defined on $\mathbb{R}^+$ is

$$F(s) = \mathscr{M}\{f(x)\}(s) = \int_0^\infty x^{s-1} f(x)\, dx, \tag{5.27}$$

where $s \in \mathbb{C}$ is a complex transform variable. Under certain restrictions on $f(x)$, $F(s)$ will be analytic in a strip parallel to the imaginary axis. The common interpretation of the Mellin transform as a Laplace transform computed on logarithmic scale is explained if we rewrite Eq. (5.27) as

$$F(s) = \int_0^\infty \exp(s \ln x) f(x) \frac{dx}{x} = \int_0^\infty \exp(sy) f(e^y)\, dy\,, \tag{5.28}$$

by virtue of the substitution $y = \ln x$.

A complex matrix-variate Mellin transform was introduced by Mathai [23], who referred to it as the M-transform. Let $f(\mathbf{X})$ be a real-valued scalar function defined on a cone Σ_+ of complex, positive definite and Hermitian matrices with dimension $d \times d$. Further let the function f be symmetric in the sense that $f(\mathbf{XY}) = f(\mathbf{YX})$, where $\mathbf{X}, \mathbf{Y} \in \Sigma_+$. The Mellin transform of $f(\mathbf{X})$ is then given by

$$F(s) = \mathscr{M}\{f(\mathbf{X})\}(s) = \int_{\Omega_+} |\mathbf{X}|^{s-d} f(\mathbf{X})\, d\mathbf{X}\,. \tag{5.29}$$

We can show that this is essentially a Laplace transform of the scalar $\ln|\mathbf{X}|$.

An important difference between the univariate and the matrix-variate Mellin transform is that the former is bijective and associated with and inverse transform, while the latter is surjective (or onto) and cannot be inverted.

5.5.2.2 Univariate Mellin Kind Statistics

Because of the domain of the Mellin transform integral, it can be directly applied to amplitude and intensity distributions. This was exploited by Nicolas [24], who introduced the Mellin kind characteristic function of the random variable X defined on $\mathbb{R}^+$ as

$$\phi_X(s) = E\{X^{s-1}\} = \mathscr{M}\{p_X(x)\}(s) \tag{5.30}$$

by replacing the Fourier or Laplace transform with the Mellin transform in the definition of the classical characteristic function. The Maclaurin series expansion of the exponential function is used to show

$$\begin{aligned}\phi_X(s) &= \int_0^\infty \exp((s-1)\ln x)\, p_X(x)\, dx \\ &= \sum_{r=0}^{\infty} \frac{(s-1)^r}{r!} \int_0^\infty (\ln x)^r\, p_X(x)\, dx \\ &= \sum_{r=0}^{\infty} \frac{(s-1)^r}{r!} \mu_X\{X\}\,.\end{aligned} \tag{5.31}$$

This proves that $\phi_X(s)$ can be expanded in terms of rth-order logarithmic moments (or log-moments) defined as $\mu_r\{X\} = E\{(\ln X)^r\}$, provided they all exist. From (5.31), it is also seen that the log-moments can be retrieved from $\phi_X(s)$ by

$$\mu_r\{X\} = \frac{d^r}{ds^r}\phi_X(s)\bigg|_{s=1}\,. \tag{5.32}$$

The Mellin kind cumulant generating function of X is further defined as $\varphi_X(s) = \ln\phi_X(s)$. This function can be expanded as

$$\varphi_X(s) = \sum_{r=1}^{\infty} \frac{(s-1)^r}{r!} \kappa_r\{X\}\,, \tag{5.33}$$

with the coefficients $\kappa_r\{X\}$ referred to as rth-order logarithmic cumulants (or log-cumulants), provided that they all exist. This is equivalent to requiring that all log-moments exist, since $\kappa_r\{X\}$ is a polynomial in the log-moments up to the same order. The log-cumulants are retrieved from (5.33) by

$$\kappa_r\{X\} = \frac{d^r}{ds^r} \varphi_r\{X\} \bigg|_{s=1}\,. \tag{5.34}$$

The first three relations between log-moments and log-cumulants are

$$\kappa_1 = \mu_1\,, \tag{5.35}$$
$$\kappa_2 = \mu_2 - \mu_1^2\,, \tag{5.36}$$
$$\kappa_3 = \mu_3 - 3\mu_1\mu_2 + 2\mu_1^3\,. \tag{5.37}$$

5.5.2.3 Matrix-Variate Mellin Kind Statistics

The domain of the complex matrix-variate Mellin transform cooincides with the domain of the covariance and coherency matrix distributions. It can therefore be used to define a Mellin kind characteristic function for the complex matrix-variate case:

$$\phi_{\mathbf{C}}(s) = E\{|\mathbf{C}|^{s-d}\} = \mathscr{M}\{p_{\mathbf{C}}(\mathbf{C})\}(s)\,. \tag{5.38}$$

This expression can be expanded into

$$\phi_{\mathbf{C}}(s) = \sum_{r=0}^{\infty} \frac{(s-d)^r}{r!} \mu_r\{\mathbf{C}\} \tag{5.39}$$

in terms of the matrix log-moments, $\mu_r\{\mathbf{C}\} = E\{(\ln|\mathbf{C}|)^r\}$, provided they exist. The matrix log-moments are retrieved from

$$\mu_r\{\mathbf{C}\} = \frac{d^r}{ds^r} \phi_{\mathbf{C}}(s) \bigg|_{s=d}\,. \tag{5.40}$$

Again we see that the Mellin transform effectively analyses data on logarithmic scale, in this case with the determinant condensing the information into a scalar.

The Mellin kind cumulant generating function in the complex matrix-variate case becomes $\varphi_{\mathbf{C}}(s) = \ln \phi_{\mathbf{C}}(s)$, whose expansion is

$$\varphi_{\mathbf{C}}(s) = \sum_{r=1}^{\infty} \frac{(s-d)^r}{r!} \kappa_r\{\mathbf{C}\} \tag{5.41}$$

in terms of the matrix log-cumulants $\kappa_r\{\mathbf{C}\}$, provided they exist. A matrix log-cumulant (MLC) can be retrieved from

$$\kappa_r\{\mathbf{C}\} = \frac{d^r}{ds^r} \varphi_{\mathbf{C}}(s) \Big|_{s=d} . \tag{5.42}$$

5.5.2.4 Application to the Product Model

The Fourier and Laplace transforms play important roles in signal processing with the additive signal model due to their convolution properties: A convolution in the input domain corresponds to a multiplication in the Fourier or Laplace transform domains. The Mellin transform has the exact same properties for the product model, as we shall see.

Assume that single-polarimetric multilook intensity can be modelled as $Y = \tau \cdot X$, i.e., a product of a texture variable τ and the Gamma distributed speckle variable X. We then have the following relations:

$$p_Y(y) = (p_\tau \hat{*} p_X)(y) = \int_0^\infty p_\tau(t) p_X(y/t)\, dt\,, \tag{5.43}$$

$$\phi_Y(s) = \phi_\tau(s) \cdot \phi_X(s)\,, \tag{5.44}$$

$$\varphi_Y(s) = \varphi_\tau(s) + \varphi_X(s)\,, \tag{5.45}$$

$$\kappa_r\{Y\} = \kappa_r\{\tau\} + \kappa_r\{X\}\,. \tag{5.46}$$

The pdf $p_Y(y)$ is computed with a multiplicative convolution, as defined in (5.43), which is the same operation we use to arrive at Eqs. (5.7) and (5.10). The multiplicative convolution operation is here denoted with the $\hat{*}$ operator. This translates to the factorisation of $\phi_Y(s)$ shown in (5.44), which carries over to an additive decomposition of $\varphi_Y(s)$ in (5.45). This leads directly to the desired property of the log-cumulants, i.e., the additive decomposition of the texture and speckle contributions in (5.46).

By using results on the matrix-variate Mellin transform from [25], we find that the product model for the polarimetric sample covariance matrix decomposes in the same manner in the different domains:

$$p_{\mathbf{C}}(\mathbf{C}) = (p_\tau \hat{\circledast} p_{\mathbf{W}})(\mathbf{C}) = \int_0^\infty p_\tau(t)\, p_{\mathbf{W}}(\mathbf{C}/t)\; dt\,, \tag{5.47}$$

$$\phi_{\mathbf{C}}(s) = \phi_\tau(d(s-d)+1)\cdot \phi_{\mathbf{W}}(s)\,, \tag{5.48}$$

$$\varphi_{\mathbf{C}}(s) = \varphi_\tau(d(s-d)+1) + \varphi_{\mathbf{W}}(s)\,, \tag{5.49}$$

$$\kappa_r\{\mathbf{C}\} = d^r \kappa_r\{\tau\} + \kappa_r\{\mathbf{W}\}\,. \tag{5.50}$$

The details of the derivations are omitted to focus on the applicative value of these results.

We particularly want to make use of the log-cumulant decomposition in (5.50), since it can be exploited to isolate speckle parameters from texture parameters. We assume for the time being that the distribution of τ is unknown, and invoke the following result for the speckle matrix $\mathbf{W}$.

The MLCs of a scaled complex Wishart distributed random matrix are known as [25]

$$\kappa_r\{\mathbf{W}\} = \begin{cases} \psi_d^{(0)}(L) + \ln|\boldsymbol{\Sigma}| - d\ln L & r = 1\,, \\ \psi_d^{(r-1)}(L) & r > 1\,, \end{cases} \tag{5.51}$$

where $\psi_d^{(r)}(x)$ is the multivariate polygamma function, defined in [25] as

$$\psi_d^{(r)}(x) = \sum_{i=0}^{d-1} \psi^{(r)}(x-i)\,, \tag{5.52}$$

and where

$$\psi^{(r)}(x) = \frac{d^{r+1}}{dx^{r+1}} \ln\Gamma(x),\ r \geq 0 \tag{5.53}$$

is the ordinary polygamma function of order r, which is an $(r+1)$th order derivative of the logarithm of Euler's Gamma function.

When including the texture contribution, the general MLCs of $\mathbf{C}$ become

$$\kappa_r\{\mathbf{C}\} = \begin{cases} \psi_d^{(0)}(L) + \ln|\boldsymbol{\Sigma}| - d(\ln L - \kappa_1\{T\}) & r = 1\,, \\ \psi_d^{(r-1)}(L) + d^r \kappa_r\{\tau\} & r > 1\,. \end{cases} \tag{5.54}$$

5.5.3 Shape Parameter Estimation

From the MLCs of $\mathbf{C}$ in (5.54), it should be observed that only $\kappa_1\{\mathbf{C}\}$ depends on the population covariance $\boldsymbol{\Sigma}$. The higher-order MLCs, $\kappa_{r>1}\{\mathbf{C}\}$, are invariant to scaling and only depend on the shape parameters. This is useful when we want to estimate the shape parameter L of the speckle variable $\mathbf{W}$ or the shape parameter(s) $\boldsymbol{\theta}$ of the texture variable τ. We also observe that the shape parameters of texture and speckle are decoupled, unlike the relations we get with linear moments. We shall take advantage of this in the sequel.

5.5.3.1 Estimation of the Equivalent Number of Looks

The shape parameter L is known in SAR literature as the equivalent number of looks (ENL). It is a parameter of multilook SAR images which describes the degree of averaging applied to the SAR measurements during data formation and postprocessing. Multilooking is performed in order to mitigate the noiselike effect of interference resulting from coherent addition of a large number of radar echoes, and sometimes also to reduce the data volume. In this process, correlated measurements are averaged, which complicates statistical modelling of the resulting multilook data. The pragmatic solution is to model the output as an average of independent measurements and to replace the actual number of correlated samples by an equivalent number of independent ones, i.e., the ENL.

We take as the ENL estimate the parameter value which produces a best match between empirical moments of correlated data and theoretical moments of the data model, which assumes independence. The ENL is therefore generally a noninteger number.

5.5.3.2 Coefficient-of-Variation Estimator

Multilook intensity of fully developed speckle is known to follow a Gamma distribution [26], and the moments of a Gamma random variable X are given by

$$E\{X^r\} = \frac{\Gamma(L+r)}{\Gamma(L)} \left(\frac{\sigma}{L}\right)^r , \tag{5.55}$$

where $\sigma = E\{X\}$ is the mean intensity. The ENL is in the literature sometimes defined as

$$L \triangleq \frac{E\{X\}^2}{Var\{X\}}, \tag{5.56}$$

since this particular combination of moments, which is known as the coefficient of variation (CV), evaluates exactly to L for Gamma distributed intensity data. However,

this definition is found to be inadequate, both because its validity is restricted to Gamma distributed intensity data and also since the given combination of moments is not the only one which is equal to L.

An alternative definition could be to say that the ENL is the shape parameter associated with a random variable representing fully developed speckle, which must be solved for (from a chosen moment expression) under the assumed statistical model, also considering the texture which may be incorporated explicitly or implicitly into this model. Such a definition allows the ENL to be a distribution parameter for different data formats, including those of polarimetric SAR.

The ENL has traditionally been estimated with samples collected from manually selected windows where the radar reflectivity is assumed to be homogeneous and the speckle fully developed, such that the sample can be assumed to follow a textureless distribution. In the following, we will accordingly assume that the samples used in the ENL estimation represent Gamma distributed intensity or scaled complex Wishart distributed sample covariance or coherency matrices.

5.5.3.3 Trace Moment Estimator

The expression in (5.56) can only be used to estimate the ENL for single-polarimetric intensity data, but has also been used with PolSAR data by computing channel-specific estimates and averaging them. We call this the coefficient of variation estimator (CVE). A weakness of this type of approach is that it does not utilise the complete information of the sample covariance matrices, only the intensities on the diagonal.

As a remedy, a generalisation of the coefficient of variation to polarimetric data was proposed in [27]. It builds upon moments of $\mathbf{C}$ involving the trace operator, that have been derived in [28]:

$$E\{tr(\mathbf{CC})\} = tr(\boldsymbol{\Sigma}\boldsymbol{\Sigma}) + tr(\boldsymbol{\Sigma})^2/L\,, \tag{5.57}$$

$$E\{tr(\mathbf{C})^2\} = tr(\boldsymbol{\Sigma})^2 + tr(\boldsymbol{\Sigma}\boldsymbol{\Sigma})/L\,. \tag{5.58}$$

These can be used to show that

$$\frac{tr(\boldsymbol{\Sigma})^2}{E\{tr(\mathbf{CC})\} - tr(\boldsymbol{\Sigma}\boldsymbol{\Sigma})} = L \tag{5.59}$$

under the scaled complex Wishart distribution. This expression is used to form the trace moment estimator (TME) [27]

$$\hat{L}_{TME} = \frac{tr(\langle\mathbf{C}\rangle)^2}{\langle tr(\mathbf{CC})\rangle - tr(\langle\mathbf{C}\rangle\langle\mathbf{C}\rangle)}\,, \tag{5.60}$$

where the operator $\langle\cdot\rangle$ denotes an average over the estimation sample.

This also serves as an example of how moment-based estimators are constructed, namely by replacing population moments with sample moments or, equivalently, by exchanging the expectation operator $E\{\cdot\}$ with the sample average $\langle\cdot\rangle$.

5.5.3.4 Maximum Likelihood Estimator

The performance of the TME is much improved with respect to the single-channel CVE. It is also very computable with its attractive closed form solution. Still, superior performance is obtained by formulating an ENL estimator based on MLCs. One can clearly solve for the ENL from MLCs of different order, but simulations have proven that the first-order MLC equation produces the best result. By assuming no texture, this equation becomes

$$E\{\ln|\mathbf{C}|\} = \ln|\boldsymbol{\Sigma}| + \psi_d^{(0)}(L) - d\ln L\,. \tag{5.61}$$

This equation does not have an analytic solution and must be solved numerically, which is easily implemented with a root finding algorithm. Notably, it is found that this equation defines the MLE under the scaled complex Wishart distribution.

In order to turn (5.61) into an estimator, both $\ln|\boldsymbol{\Sigma}|$ and $\boldsymbol{\Sigma}$ must be replaced by empirical averages. With this in place, the MLE estimate $\hat{L}_{MLE}$ is defined as the root of

$$g(\hat{L}_{MLE}) = \langle\ln|\mathbf{C}|\rangle - \ln|\langle\mathbf{C}\rangle| - \psi_d^{(0)}(\hat{L}_{MLE}) - d\ln\hat{L}_{MLE}\,. \tag{5.62}$$

Although the population covariance is a nuisance parameter, the MLE performs very well in terms of its mean square error, as shown in [27].

The ENL value is an image constant which is determined by the processing scheme chosen for the data set. It should therefore be estimated prior to any other distribution parameter. The ENL value will change with the filters and processing parameters used in the frequency domain multilooking performed during SAR focusing or in spatial domain postprocessing of the focused image. It is therefore important to have a robust and accurate algorithm for ENL estimation. An automatic algorithm is also preferred, which avoids manual selection of the estimation sample. An unsupervised algorithm which extracts the overall ENL estimate from small sample estimates computed in local windows across the image is described in [27]. It is inspired by a previous attempts at a similar procedure from [29].

The estimators presented above all assume that the estimation sample is homogeneous and represents fully developed speckle without any textural influence. The associated unsupervised estimation procedure relies on the assumption that this holds for a sufficient number of the windows where small sample estimates are computed,

such that estimates produced under ideal circumstances will dominate the collection of estimates. It is evident that occurrence of texture, class mixtures and other departures from the model assumption will bias the overall estimate. As a resort, a texture invariant estimation procedure has been proposed in [30], whereas a mixture eliminating procedure is presented in [31].

5.5.4 Estimation of Texture Parameters

We shall now demonstrate how the MLCs can be used to estimate texture parameters. This will be done for a Gamma distributed texture variable with one shape parameter and for a Fisher distributed texture variable with two shape parameters. These texture distributions are shown in Table 5.1, where both are normalised to unit mean.

The univariate log-cumulants of a Gamma distributed τ with shape parameter α are

$$\kappa_r\{\tau\} = \begin{cases} \psi^{(0)}(\alpha) - \ln\alpha & r = 1\,, \\ \psi^{(r-1)}(\alpha) & r > 1\,. \end{cases} \tag{5.63}$$

For a Fisher distributed τ with shape parameters α and λ, the univariate log-cumulants are

$$\kappa_r\{\tau\} = \begin{cases} \psi^{(0)}(\alpha) - \psi^{(0)}(\lambda) + \ln\left(\frac{\lambda-1}{\alpha}\right) & r = 1\,, \\ \psi^{(r-1)}(\alpha) + (-1)^r\psi^{r-1}(\lambda) & r > 1\,. \end{cases} \tag{5.64}$$

The texture parameters of both these models can be estimated by the method of matrix log-cumulants (MoMLC). In this method we set up as many MLC equations as we have unknown distribution parameters. We then replace the population MLCs with corresponding sample MLCs and solve for the unknown parameters. This must be done numerically.

The Gamma distribution contains only one shape parameter and therefore requires only one MLC equation. The Fisher distribution requires two. Low-order MLCs are preferred, since the estimation variance generally increases with moment order, although this also depends on the presence of nuisance parameters. For instance, the first-order MLC contains the population covariance, but can still be used in the MoMLC approach, both alone and together with other MLC expressions.

We have also proposed another estimation method which utilises more moment orders than the MoMLC in order to capture more information about the unknown parameters. This yields an overdetermined system of nonlinear equations. Let $\boldsymbol{\kappa}$ be a vector of distinct MLCs with selected order and $\langle\boldsymbol{\kappa}\rangle$ a vector containing certain sample statistics of corresponding order. Specifically, the entries of $\langle\boldsymbol{\kappa}\rangle$ must be a version of the so-called k-statistics [32]. These are minimum-variance unbiased estimators of the MLCs, to ensure that unbiasedness is fulfilled in terms of: $E\{\langle\boldsymbol{\kappa}\rangle\} = \boldsymbol{\kappa}$.

We must further use the covariance matrix $\mathbf{K} = E\{(\langle\boldsymbol{\kappa}\rangle - \boldsymbol{\kappa})(\langle\boldsymbol{\kappa}\rangle - \boldsymbol{\kappa})^T\}$, which can be derived from the definition of the k-statistics and cross-covariance relations between sample moments, all given in [32]. Knowing that $\boldsymbol{\kappa}$ depends on the texture parameters through τ, we define the maximum asymptotic likelihood (MAL) estimator:

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \left\{ n(\langle\boldsymbol{\kappa}\rangle - \boldsymbol{\kappa})^T \mathbf{K}^{-1} (\langle\boldsymbol{\kappa}\rangle - \boldsymbol{\kappa}) \right\} . \tag{5.65}$$

The efficiency of the MAL estimator is shown in [33].

5.6 Image Classification

The previous sections have detailed the potentially non-Gaussian statistical distributions that result from the polarimetric product model and the MoMLC estimation techniques used to estimate their model parameters. We will now explain how to use both of these concepts for non-Gaussian image classification. This is first done for supervised classification into known classes with training data, and secondly for unsupervised segmentation (clustering) into unknown segments without training data.

We will study the case of multilook complex polarimetric covariance matrix data assuming an underlying Fisher distributed texture parameter and thus a matrix-variate U-distribution as model for the compound sample covariance matrix, as it is the most flexible. The matrix-variate K-distribution and the complex Wishart distribution will be special cases of this model. Be aware that the statistical methods described here could equally well be applied to other models and data, including single-look complex vector data, with the appropriate choice of probability density function.

Both of these cases make rigorous statistical use of the model probability density functions, which is essentially adding model specific information and constraints into the decision-making process. This is distinctly different from non-model-based methods, such as support vector machines (supervised, with training data only), non-parametric or distance-based methods, that do not utilise or impose any model information. Further discussion of these alternative methods is outside the scope of this chapter.

5.6.1 Supervised Classification

For the case of supervised classification with the matrix-variate U-distribution, we assume that each pixel $\mathbf{C}_i$ belongs to a class defined by the distribution $U(\mathbf{C}; L, \Sigma, \alpha, \lambda)$ from Table 5.2. We also have some amount of known training data samples that can be used to define the different classes.

Firstly, the parameters for each U-distributed class need to be estimated from the data samples for that class. Let us define the N_j training data samples for class j as $\{\mathbf{C}_{T_j(k)} : k = 1, \ldots, N_j\}$. Then the estimates for the class conditional parameters are determined by

$$\hat{\Sigma}_j = \frac{1}{N_j} \sum_{k=1}^{N_j} \mathbf{C}_{T_j(k)} \tag{5.66}$$

$$\hat{\alpha}_j, \hat{\lambda}_j = \underset{\alpha_j, \lambda_j}{argmax} \left(\boldsymbol{\kappa}_j - \langle \boldsymbol{\kappa} \rangle_j\right) \mathbf{K}_j^{-1} \left(\boldsymbol{\kappa}_j - \langle \boldsymbol{\kappa} \rangle_j\right) , \tag{5.67}$$

where $\boldsymbol{\kappa}_j$ is a vector of population log-cumulants that follow from the model, $\langle \boldsymbol{\kappa} \rangle_j$ is the vector of corresponding k-statistics, which provides an unbiased estimate of $\boldsymbol{\kappa}_j$, and $\mathbf{K}_j$ is the covariance matrix of $\langle \boldsymbol{\kappa} \rangle_j$. Both $\boldsymbol{\kappa}_j$ and $\mathbf{K}_j$ are functions of the parameters α_j and λ_j, while $\langle \boldsymbol{\kappa} \rangle_j$ depends on the data sample. This method of (5.67) was introduced in Sect. 5.5.4 as the MAL estimator [33].

In addition, we must decide whether we wish to include any class prior probability values (also known as mixing proportions), $\pi_j : j = 1, \ldots, J$, or take the non-informative, equiprobable case, $\pi_j = 1/J$. The former determines the *maximum a posteriori* probabilities, whilst the latter determines the *maximum likelihood* probabilities. The choice depends on whether you know that the classes ought to have different overall abundances in the image and that you can reflect that knowledge in the prior probability weights. If you do not know, then you should take the safer maximum likelihood approach. In either case, Bayesian decision theory says that you should label each pixel $\mathbf{C}_i$ to the class with the greatest posterior likelihood, that is, assign the pixel to the class j such that

$$U(C_i|L, \Sigma_j, \alpha_j, \lambda_j)\pi_j \quad > \quad U(C_i|L, \Sigma_k, \alpha_k, \lambda_k)\pi_k \qquad \forall\, k \neq j\,. \tag{5.68}$$

Since the model includes the non-Gaussian texture parameters for each class distribution, then the decision is based upon non-Gaussian modelling and includes more intensity variation for any highly textured classes, and less variation for the more homogeneous classes.

It is important to realise some of the assumptions of this method. The probabilistic decision will always find the single most-likely class out of those classes present. Therefore, if you do not have fully complete training data, and are actually missing some important classes, then the pixels will be assigned to the most similar class of those present. To avoid this you would have to represent all possible classes, or to make a sort of residual class with some specific probabilistic threshold. If the probability is greatest for class j and it is above a certain threshold, then assign the pixel to that class, otherwise assign it to the "unknown" or "uncertain" residual class.

The number of looks is an important parameter for non-Gaussian probability density functions and does not cancel away in the relation in (5.68), as in the case for the purely Gaussian/Wishart model. Hence, the equivalent number of looks must

be estimated for the modelling and several approaches to do this are outlined in Sect. 5.5.3.1. Most important for all of these non-Gaussian models is the concept of class mixtures. The estimation routines based on the modelling theory will not be appropriate if your samples are not from one uniform class with common parameters, but from mixtures of several classes. Mixtures will have more variation than a uniform class, since you get the speckle and texture variation, plus you get some component of variation relating to the differences between the classes. This is often mistaken for textural variation, and often leads to extreme degrees of texture in the modelling. This concept of mixtures is compounded because the log-cumulant and moment based methods are power based and will be severely influenced by outliers from class mixtures. Furthermore, the probability density functions for extreme texture tend to be very broad and likelihood based decisions are overly accepting of virtually all data samples leading to poor class distinction. This sensitivity may be mitigated by either taking care to have enough classes with pure uniform samples, or by suppressing a small fraction of outliers before estimating the class parameters. The latter is an easy way to improve the model fitting and still benefit from the fast and simple moment or cumulant based estimation methods. Alternatively, you would need a numerical approach that emphasises the peak values and is not so sensitive to outliers in the tails, e.g., a numerical maximum-likelihood method.

For a data example, we will look at a RADARSAT-2 sample SAR scene from Vancouver, Canada (available for free download from the MDA web-site [34]). This scene includes some water, urban and vegetated areas in and around the city of Vancouver, Canada. The original single-look complex data-set was spatially multi-looked with a large window, 22×11 (azimuth $\times$ range), and strongly sub-sampled to simplify the image segmentation into only a few regional (large-scale) land-cover classes. To obtain training data, we manually selected polygon regions for five major classes: (1) City centre; (2) Forest; (3) Agricultural fields; (4) Ocean; and (5) Urban. These regions may be seen in the Pauli RGB representation in Fig. 5.1.

The supervised classification estimates the class parameters from the training region with the ENL and texture parameters estimated by optimisation of an overdetermined system of MLC equations, as explained in Sect. 5.5.4, i.e., by using the maximum asymptotic likelihood estimator from [33]. Note that we included a probabilistic outlier removal of 1% to reduce the influence of mixtures in the estimation. The resulting class histograms and model fitting are shown in Fig. 5.1, where this one-dimensional histogram depicts $trace(\boldsymbol{\Sigma}^{-1}\mathbf{C})$ which compacts the matrix-variate data into a simple visualisation related to the texture distribution and the different widths indicate the degree of texture in each class. The city class (1, red) has the most texture, while the forest (2, green) has the least textural variation in this case.

The supervised maximum likelihood classification result is shown in Fig. 5.2, and it is quite apparent that it separates the chosen training classes quite well, when compared visually to the Pauli RGB image. There is some confusion in the upper mountainous regions, but we have not applied any form of terrain incidence angle correction to this highly sloping region, hence the terrain-slope brightness variations are appearing as several different classes. One may also observe that some areas like the darker runways at the airport are being classified to the nearest available class of

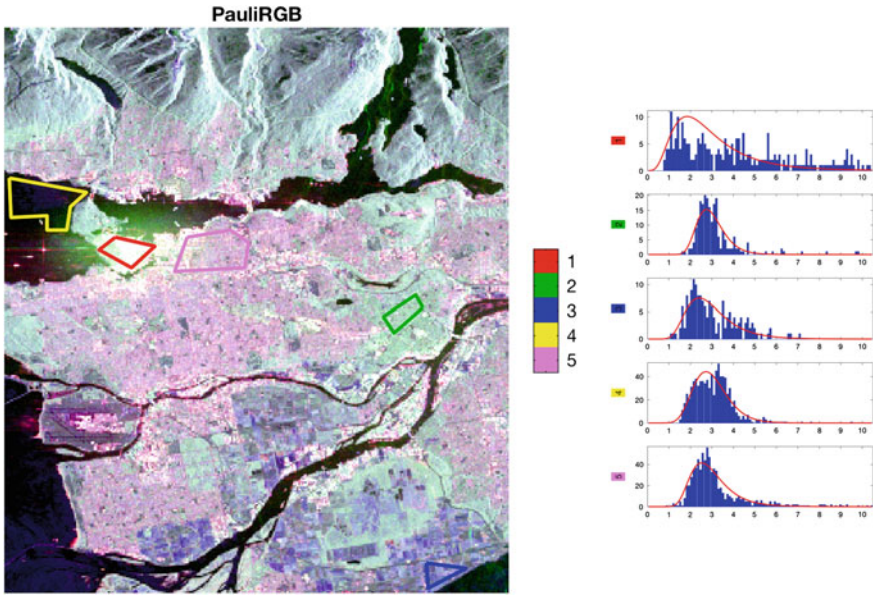


Fig. 5.1 Pauli RGB image (R = HH − VV, G = HV, B = HH + VV), with five coloured training regions marked, and the corresponding training class histograms with the fitted model curve in *red* (color figure online)

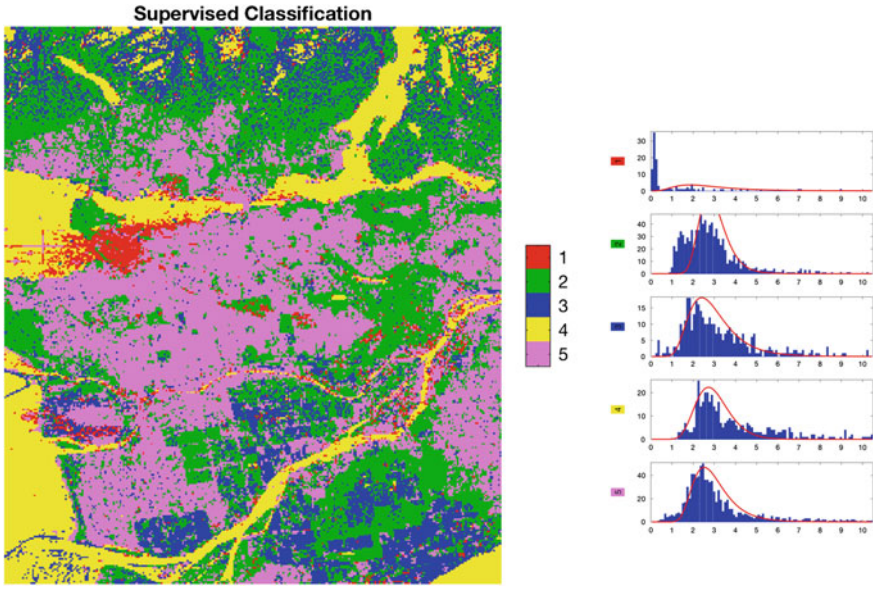


Fig. 5.2 Supervised classification result for the five coloured training regions and their resulting class histograms. Note that some mixed histograms now appear, due to too few classes

the five supervised choices. The resulting class histograms show some clear mixing with evident multiple peaks and poor fitting model curves.

5.6.2 *Unsupervised Segmentation*

For the case of unsupervised segmentation, or clustering, with the matrix-variate U-distribution, we assume as before that each pixel $\mathbf{C}_i$ belongs to a class following the U-distribution $U(\mathbf{C};\, L,\, \Sigma, \alpha, \lambda)$ from Table 5.2. Furthermore, the total probability model for all J classes in the image is a finite mixture model described by

$$p(\mathbf{C}) = \sum_{j=1}^{J} \pi_j \; U(\mathbf{C};\, L,\, \Sigma_j, \alpha_j, \lambda_j) \tag{5.69}$$

where each class j has a prior probability π_j and class conditional parameters $\Sigma_j, \alpha_j, \lambda_j$. This type of problem is known as *finite mixture modelling* and has many hidden class parameters. It is generally solved with an expectation maximisation algorithm (EM-algorithm) [35, 36]. The EM-algorithm is an iterative algorithm that repeatedly calculates the expected posterior likelihood, given the data and the current parameter values, and then updates the current parameter values given the class posterior likelihoods. It must be initialised with some starting value (often randomly assigned) and then iterates until some convergence criterion is met (often the change in total log-likelihood). The end result is then the estimated set of all class parameters and their mixing priors (or proportions). The posterior membership likelihoods may also be extracted from the final iteration, or recalculated based on the class parameters for all samples of interest and for each class. These posterior memberships can then be used to make a Bayesian decision like the supervised case from (5.68), but where the class parameters have been estimated automatically.

Note that although the resulting 'classes' have statistically distinguishable properties, they are essentially unknown 'classes', with a randomly assigned index label. We often try to avoid using the word 'classes' here, and instead use the words clusters or segments, to indicate this missing information. A true classification, into known and named classes, may be subsequently made using auxiliary knowledge, like limited training data or class physical properties.

This approach too has certain assumptions and limitations that affect the interpretation and quality of the result, and yet it will always return a segmented image. It assumes that the pdf model is a good shape for the real data, as this will guide where it determines class boundaries. It requires the number of clusters to be set in advance, and much like the supervised method needing representative samples, a wrong number of classes will create some mixed clusters or some randomly split clusters. The effect of this is often quite visible, and one practically useful strategy is to perform unsupervised clustering to several different numbers of clusters, and then pick the best looking image result visually. This sequential procedure has even

been automated, by the chapter authors, by incorporating a goodness-of-fit test to judge the appropriate number of clusters [37]. This is only achievable because the model probability density function information gives us extra shape information. Furthermore, the EM-algorithm is influenced by the choice of initial conditions and random initialisation can lead to many different resulting images. The usual tricks of performing multiple starts and picking the greatest likelihood score, or seeding with a semi-supervised or simpler pre-classified result can help regularise the resulting segmentation.

As an example for unsupervised segmentation, we will take the same image as the supervised case and firstly set the number of clusters also to five, and subsequently to a higher number of 9 classes. This will show some of the aspects to consider with such clustering.

Firstly, if the number of classes is too few compared to the actual number of distinctly different regions, then some mixed clusters must occur, in a similar way as too few training classes in the supervised approach. However, for unsupervised, you have no control over which clusters the algorithms mixes and it may not be the same ones as your supervised choice. This has occurred in our five class example in Fig. 5.3, where the blue training class 3 is not distinguished, and a new area of urban is found instead. A visual check with the PauliRGB (Fig. 5.1) and we can easily see that this new region is in fact representing a real area of much brighter pink than the rest of the urban, but it doesn't match our manual choice. Additionally, the

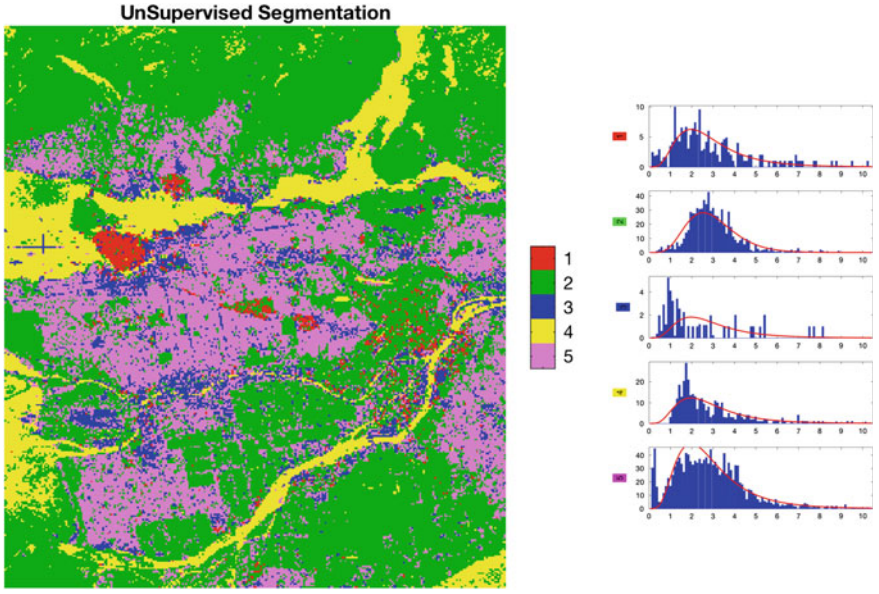


Fig. 5.3 Unsupervised segmentation into 5 classes and histograms

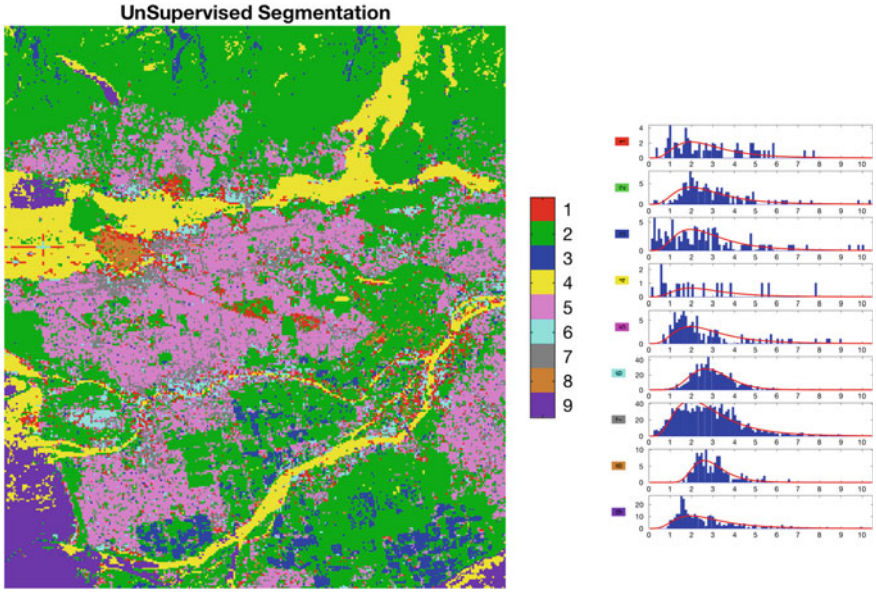


Fig. 5.4 Unsupervised segmentation into 9 classes and histograms

histograms indicate that class 2 (green) and class 5 (pink) appear to be significantly mixed, as the model curve does not match and there are even multiple peaks.

Secondly, we can break this mixed condition by allowing several more clusters to adapt to these real distinct regions, thus reducing the mixed clusters and achieving better cluster histogram fitting and better visual results. Figure 5.4 shows an unsupervised segmentation for nine clusters and their histograms. See how the histograms better match the fitted model curves and that the distinguished regions can be visually compared to different coloured regions in the Pauli RGB image of Fig. 5.1.

Finally, note that the clusters seem to have better matched the colour boundaries in the Pauli RGB image than the unsupervised case. This is because the extra clusters can fit tighter around the actual data histograms than the slightly mixed classes from the training data polygons. If desired, the number of classes can be reduced after segmentation by combining particular segments, for example the purple and yellow water sub-classes.

These examples only scratch the surface of the possibilities for model-based non-Gaussian image analysis and are only intended to reflect some of the major considerations and to demonstrate that it achieves reasonable results.

5.7 Coherent Time-Frequency Characterization of Complex Polarimetric Features

The second part of this chapter goes deeper into the multi-pulsed synthetic acquisition of a the PolSAR signal and explores sub-aperture consistency. Conventional SAR image analysis and geophysical parameter retrieval techniques from SAR data generally assume that scenes are formed of static scatterers observed in the direction perpendicular to the flight track and at a fixed frequency, equal to the emitted signal carrier's one. When imaging complex objects and media, potential variations of the signal measured during the SAR acquisition may strongly affect feature estimates derived from the resulting SAR data and may lead to erroneous interpretations. This kind of phenomenon, frequently encountered with moving objects [38], may be observed with static objects with anisotropic geometrical structures or having a frequency selective response. Their electromagnetic behavior may vary as they are illuminated from different positions or at different frequency components during SAR integration. The resulting SAR response being well described by the spatial convolution of a conventional scene SAR image with specific functions accounting for each effect, non ideal features can be easily detected and characterized in the spectral domain.

This section presents different techniques for detecting scatterers having a varying response during the SAR acquisition and characterizing the underlying physical phenomenon that generates these variations. These approaches are based on specific coherent Time-Frequency (TF) decompositions which can be applied on already focused SAR images.

5.7.1 Coherent Time-Frequency Decomposition of Polarimetric SAR Images

As depicted in Fig. 5.5, and explained in this chapter, a SAR measurement consists in repeatedly emitting a signal, $s_e(t)$, in the across track direction and receiving the echo from the observed scene, $s_r(x,t)$, at different locations x along the acquisition track. A scatterer P_0 located at coordinates $(x_0, y_0, 0)$ is observed for different values of the azimuth look angle, ϕ, defined by $\sin(\phi) = (x - x_0)/d_0(x)$, defined $d_0(x) = \sqrt{r_0^2 + (x - x_0)^2}$ being the varying radar-scatterer distance. The range of the azimuth angle is defined by the antenna aperture, whereas the emitted signal is characterized by a bandpass spectrum centered around a carrier frequency f_c, with bandwidth B_f.

Under simplifying assumptions and considering ideal acquisition conditions [39], a coherent SAR image may be formulated as a convolution of the scene coherent reflectivity, $\gamma(x,r)$ and the SAR 2-D impulse response, $h(x,r)$ and may be represented in both spatial and spectral domains as:

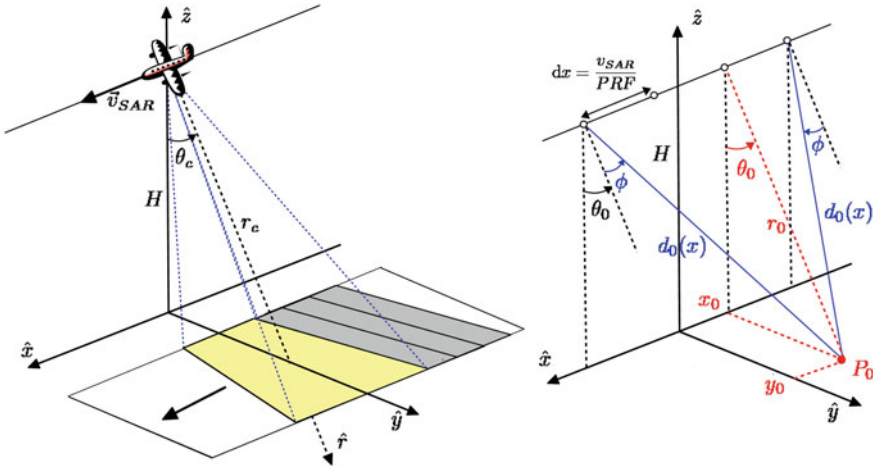


Fig. 5.5 Geometrical configuration of a SAR acquisition. Repeated emission-reception of signals, with a rectangular antenna aperture (*left*). Observation of a scatterer P_0 from different positions along the acquisition track, with corresponding azimuth look angles

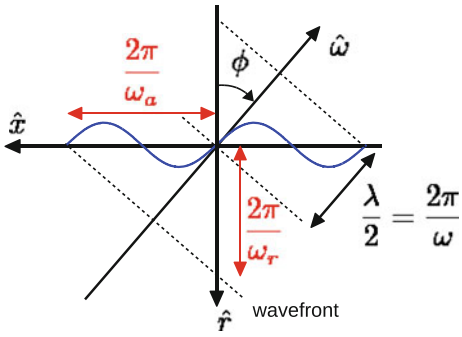


Fig. 5.6 Decomposition of a plane wave propagating along $\hat{\omega}$ into azimuth and range components. A spherical wave may be represented as a sum of plane waves with varying azimuth orientation ϕ

$$s(x,r) = \gamma(x,r) \otimes h(x,r) \equiv S(\omega_a,\omega_r) = \Gamma(\omega_a,\omega_r)\, H(\omega_a,\omega_r) \tag{5.70}$$

where $\equiv$ denotes logical equivalence, $S(\omega_a,\omega_r) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} s(x,r)\mathrm{e}^{-j\omega_a x}\mathrm{e}^{-j\omega_r r}\mathrm{d}x\mathrm{d}r$ is the 2-D Fourier transform of the SAR image $s(x,r)$, and $\otimes$ represents the convolution operator. The spectral coordinates (ω_a,ω_r), representing two-way wavenumbers in range and azimuth, are illustrated in Fig. 5.6 and can be formulated as [40], $\omega_a = \omega\sin\phi$ and $\omega_r = \omega\cos\phi - \omega_c$. where ω is the emitted signal wavenumber and can be related to the electrical frequency using, $f \in f_c + \left[-\frac{B_f}{2}, \frac{B_f}{2}\right]$, through the wave propagation velocity, c, as $\omega = \frac{4\pi f}{c}$.

For an ideal scene, whose reflectivity is uniformly distributed over the spectral domain, the resolutions of the SAR image are driven by the impulse response of the

SAR system and are given by

$$\delta x = \frac{2\pi}{\Delta\omega_a} = \frac{c}{4 f_c \sin(\Delta\phi/2)} \quad \text{and} \quad \delta r = \frac{2\pi}{\Delta\omega_r} = \frac{c}{2B_f} \tag{5.71}$$

where $\Delta\phi$ stands for the processed azimuthal aperture, whose maximal value is set by the acquisition antenna characteristics.

The T-F decompositions technique selected here is based on the use of a 2-D windowed Fourier transform, or 2-D Gabor transform. This kind of transformation permits to decompose a two-dimensional signal, $s(\mathbf{l})$, with $\mathbf{l} = [x, y]^T$ a 2-D location, into different spectral components, using a convolution with an analyzing function $g(\mathbf{l})$, as follows [41]:

$$s(\mathbf{l}_0; \boldsymbol{\omega}_0) = \int s(\mathbf{l})\, g(\mathbf{l}_0 - \mathbf{l})\, \exp(-j\boldsymbol{\omega}_0^T(\mathbf{l}_0 - \mathbf{l}))\mathrm{d}\mathbf{l} \tag{5.72}$$

where $\boldsymbol{\omega}_0 = [\omega_x, \omega_y]^T$ indicates a position in frequency, and $s(\mathbf{l}_0; \boldsymbol{\omega}_0)$ represents the decomposition result around the spatial and frequency locations $\mathbf{l}_0$ and $\boldsymbol{\omega}_0$. The application of a Fourier transform to (5.72) shows that the spectrum of $s(\mathbf{l}_0; \boldsymbol{\omega}_0)$ is given by the product of the original signal spectrum and the transform of the analyzing function g shifted around the frequency vector $\boldsymbol{\omega}_0$:

$$S(\boldsymbol{\omega}; \boldsymbol{\omega}_0) = S(\boldsymbol{\omega})\, G(\boldsymbol{\omega} - \boldsymbol{\omega}_0) \tag{5.73}$$

It is clear from Eqs. (5.72) and (5.73) that this time-frequency approach may be used to characterize, in the spatial domain, behaviors corresponding to particular spectral components of the signal under analysis, selected by the analyzing function g. Among the wide variety of existing TF analysis methods, the simple atomic decomposition selected in this study presents some interesting properties. It is linear, and hence preserves the coherence and energy of signals, it is not affected by artifacts related to cross-terms and may be inverted, i.e., depending on the analyzing function g, $s(\mathbf{l})$ may be reconstructed from a set of TF samples $s(\mathbf{l}; \boldsymbol{\omega}_0)$, provided that some sampling conditions in spatial and spectral domains are satisfied. The resolutions of the analysis in space and frequency are not independent, and their product is fixed by the Heisenberg–Gabor uncertainty relation, given in 1-D by [41], $\delta l\, \delta\omega = u_g$.

In practice the simple SAR image model given in (5.70) needs to be completed in order to account for additional weighting terms, mainly due to the antenna pattern and side-lobe reduction functions, as [42] $S(\boldsymbol{\omega}) = \Gamma(\boldsymbol{\omega})\, H(\boldsymbol{\omega})\, W(\boldsymbol{\omega})$ with $\boldsymbol{\omega} = [\omega_a, \omega_r]^T$. The synopsis of the TF decomposition based on the spectral definition on (5.73) is given in Fig. 5.7.

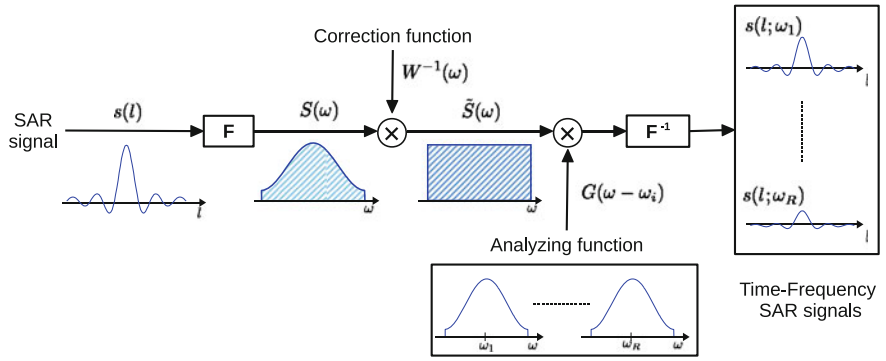


Fig. 5.7 Synopsis of the proposed Time-Frequency decomposition of a mono-dimensional SAR signal

The first step consists of correcting for potential spectral imbalances, represented by $W(\boldsymbol{\omega})$, in the original, full-resolution SAR image. This can be achieved by calculating average image spectra in range and azimuth and then multiplying the full-resolution spectrum $S(\boldsymbol{\omega})$ by the inverse of the estimated 2-D weighting function. The TF decomposition is then conducted by multiplying the corrected spectrum by the Fourier transform of the analyzing function and going back to the spatial domain. The resulting still focused SAR image $s(\mathbf{l};\boldsymbol{\omega}_0)$ has a lower resolution than the original SAR data and depicts the scene behavior over the 2D frequency domain located in the neighborhood of $\boldsymbol{\omega}_0$. In order to emphasize the physical interpretation of coherent SAR image analysis, one may simplify the wavenumber expressions given previously using narrow and bandwidth approximations

$$\omega_a \approx \omega_c \sin\phi, \quad \omega_r \approx \omega - \omega_c \tag{5.74}$$

It is worth noting that the direct relation between a SAR image and the reflectivity of scene in (5.70), as well as the physical meaning of the spectral coordinates in (5.74) are valid when dealing with coherent Single Look Complex (SLC) SAR data sets, i.e., each pixel of the SAR image corresponds to a complex number whose modulus is proportional to the focused reflectivity and whose absolute phase depends on the observed medium as well as on the measurement phase history. Transforming an SLC image to an incoherent one, like an intensity image $I(x,r) = |s(x,r)|^2$, involves an irremediable loss of information and interpretation. This kind of analysis can be applied to detect objects or media with anisotropic behaviors, like scatterers with complex geometrical structures, human-made objects, or natural media having periodic structures in the case of agricultural areas, or linear alignments of strong scatterers [42]. In the range direction, TF analysis permits to compare the response of a scene observed at different frequencies, contained within the emitted signal spectral domain, and can be used to detect and characterize media with frequency-sensitive

responses, like resonating spherical or cylindrical objects, periodic structures, or coupled scatterers with interfering characteristics [43, 44].

Polarimetric SLC SAR images can be easily decomposed by applying the presented approach independently over each polarization channel. Usual polarimetric representations may then be reconstructed to study the polarimetric behavior of a scene around specific spectral locations.

$$\mathbf{k}(\boldsymbol{\omega}_0) = \frac{1}{\sqrt{2}} \begin{bmatrix} S_{hh}(\boldsymbol{\omega}_0) + S_{vv}(\boldsymbol{\omega}_0) \\ S_{hh}(\boldsymbol{\omega}_0) - S_{vv}(\boldsymbol{\omega}_0) \\ 2S_{hv}(\boldsymbol{\omega}_0) \end{bmatrix} \quad \text{and} \quad \mathbf{T}(\boldsymbol{\omega}_0) = \mathrm{e}\left[\mathbf{k}(\boldsymbol{\omega}_0)\mathbf{k}^{\mathrm{H}}(\boldsymbol{\omega}_0)\right] \tag{5.75}$$

where spatial locations, $\mathbf{l}$, have been omitted.

5.7.2 *Characterization of Natural Environments with Non-stationary Polarimetric TF SAR Responses*

5.7.2.1 Non-stationary TF POLSAR Responses

The TF decomposition scheme is applied onto polarimetric SAR data acquired by the DLR E-SAR sensor, at L band, over the Alling test site in Germany. The original image resolution is 2 m in range and 1 m in azimuth, corresponding to an azimuthal variation of the look angle of approximately 7.5° and a chirp bandwidth of 75 MHz. Figure 5.8 shows the full-resolution span image corresponding to the total polarimetric backscattered power. The considered scene is mainly composed of agricultural fields and forest. An urban area is located at the bottom left corner of the image. The decomposition results obtained in both range and azimuth directions over an area

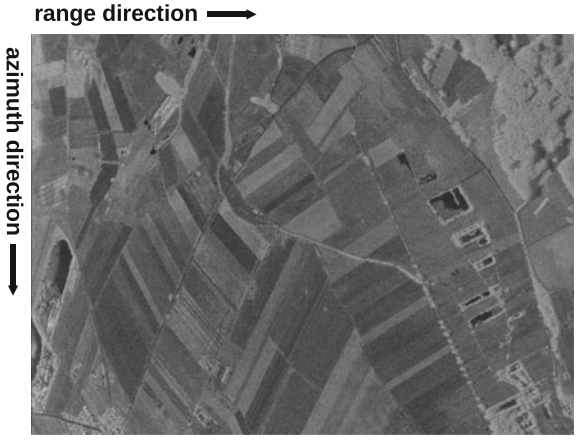


Fig. 5.8 Full resolution span image of the Alling test site

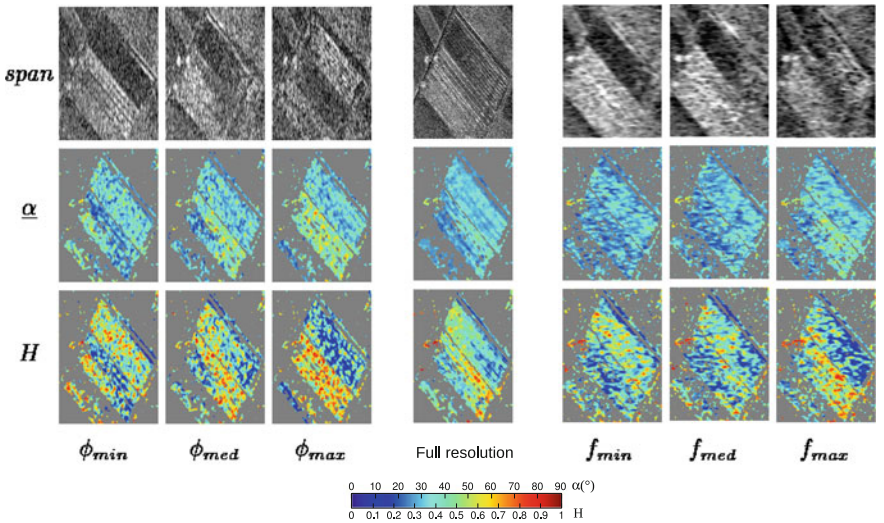


Fig. 5.9 Polarimetric parameters at full resolution (*center*) and after 1-D TF analysis in the azimuth (*left*) and range (*right*) directions

corresponding to plowed fields, are shown in Fig. 5.9, under the images formed of the span, H, and $\underline{\alpha}$ parameters derived from subspectra centered around different spectral locations and for the full-resolution case. The entropy, H, represents the degree of randomness of the polarimetric information. H equals 0 for a deterministic response and reaches 1 in case of polarimetric white noise, i.e., for uncorrelated polarimetric channels with equal intensity. The parameter $\underline{\alpha}$ is an indicator of the nature of the scattering mechanism. A value close to 0 indicates a single bounce reflection, characteristic of scattering by rough surfaces, $\underline{\alpha} = \pi/4$ corresponds to the scattering from anisotropic objects, whereas an α value close to $\pi/2$ denotes double-bounce scattering [14]. It can be observed in Fig. 5.9 that large variations in both parameters occur while the azimuth look angle or the illumination frequency change. For particular observations conditions, some fields show a sudden change of behavior: the span reaches a maximum value, whereas the polarimetric indicators H and α are characterized by low values. Such a kind behavior was found to be characteristic of Bragg resonant scattering over periodic surfaces in [42, 43].

Bragg resonance is due to the coherent summation of simultaneously constructive contributions from a set of scatterers and is likely to happen during the observation of periodic surfaces or randomly irregular surfaces with a strong periodic component, as described in 1-D in Fig. 5.10 A random surface, $h(x, y)$, with a quasi-periodic component in the y direction, can be described as

$$h(x, y) = B \cos\left(\frac{2\pi}{P}\right) + \psi(x, y) \tag{5.76}$$

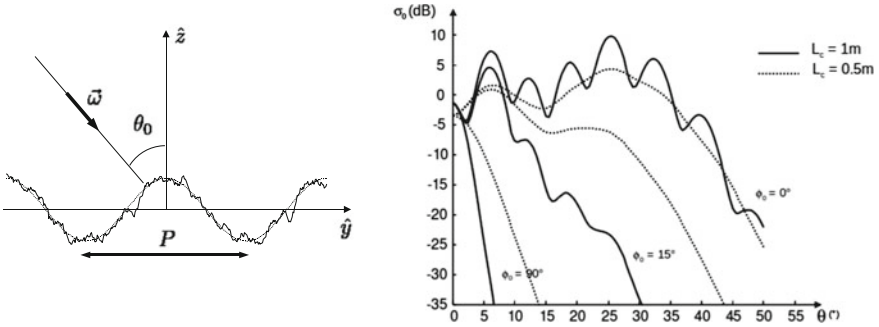


Fig. 5.10 Example of randomly perturbed periodic surface (*left*) and its associated backscattering coefficient, at L-band, with $B = 10\,\text{cm}$, $P = 1\,\text{m}$, $\varepsilon = 9$, and $\sigma_h = 1\,\text{mm}$

where P and B are the spatial period and amplitude, respectively, of the periodic component of $h(x, y)$, and the random perturbation term, $\psi(x, y)$, corresponds to an isotropic stationary random rough surface. This component is fully described by σ_h, the standard deviation of its zero mean Gaussian height probability density function, and φ_ψ, its correlation function. The Bragg resonance condition can be written as a function of the incident wavelength, λ, as

$$\omega_y = n\frac{2\pi}{P} \quad \text{or} \quad \sin\theta_0 \cos\phi_0 = \frac{n\lambda}{2P} \tag{5.77}$$

where $\omega_y = \omega \sin\theta_0$ corresponds to the local amplitude of the ground wave vector at the surface, n is an unknown integer number indicating the mode of the resonance, θ_0 is the local angle of incidence and ϕ_0 the azimuthal angular difference between the observation position and the normal to the rows of the periodic surface. In the case of SAR measurements, ϕ_0 can be decomposed as $\phi_0 = \phi_t + \phi$, where ϕ_t represents the orientation of the surface with respect to the normal to the SAR platform flight track, and ϕ with the angle of observation in the azimuthal spectrum, as defined in (5.74). Yueh [45] developed several approaches to model the scattering of electromagnetic waves from randomly perturbed periodic surfaces. Their study reports that the influence of the resonating modes on the total backscattering response varies significantly with the surface parameters. As it can be seen in Fig. 5.10, for a large surface correlation length, l_c, and for low values of the azimuth orientation angle, ϕ_0, almost all the intensity peaks corresponding to different resonance modes can be discriminated. As l_c increases, the scattering pattern becomes smoother and only a few dominant resonance peaks can be observed. In the presence of resonance, the co-polarization returns S_{HH} and S_{VV} have almost identical values, characterized by a high intensity. As the resonant effect decreases, i.e., for high values of ϕ_0, these polarimetric channels have distinct responses, with a significantly reduced amplitude. According to the resonance condition enounced in (5.77) similar anisotropic fields with different locations in range, θ_0, or differently oriented, i.e., with different

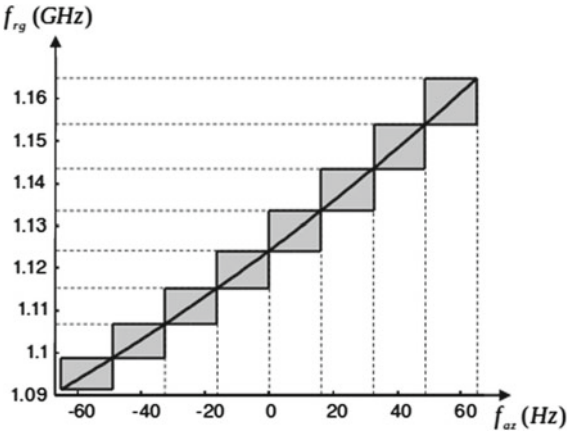


Fig. 5.11 Location of resonance peaks in the (frg, faz) plane. The *solid line* indicates the location of a resonance peak for a periodic surface characterized by P = 0.6 m and observed at L band. *Grey* areas indicate the location of potential resonance areas for each range-azimuth sub-spectrum

ϕ_t values may resonate at different azimuthal frequencies. If the resonance conditions cannot be satisfied for any azimuthal angle within the antenna aperture or if the surface scattering characteristics do not show a resonance peak, they also might not resonate at all [43].

Moreover, some fields may have parts resonating at different positions in the azimuthal frequency domain due to the joint dependence of the resonance condition on the incidence and azimuth angles, as seen in (5.77). This phenomenon is illustrated in Fig. 5.11, where the location of a resonance peak is plotted as a function of the range and azimuth frequencies. As the azimuthal look angle, ϕ, varies, the set of incidence angles θ_0 satisfying Eq. (5.77) changes, leading to the apparition of sliding resonating stripes in the (f_{rg}, f_{az}). The width of the resonating stripes is fixed by the width of the analyzing function in the azimuth and range directions, $(\Delta_g \omega_r, \Delta_g \omega_a)$ or equivalently $(\Delta_g f, \Delta_g \phi)$.

For the purpose of identifying Bragg resonance, a range-azimuth continuous time-frequency analysis is performed over three points $(P1, P2, P3)$, located at different range positions inside a plowed field [43]. As depicted in Fig. 5.12, results can be represented, for each point, in the range-azimuth frequency plane. The results of the time-frequency analysis, as shown in Fig. 5.12, demonstrate that all three points under investigation do not have a stationary range and azimuth scattering behavior. Some (ω_r, ω_a) pairs show high span values corresponding to low H and α. These observations agree with the predictions of the scattering model developed by Yueh [45]. As the surface resonates, the co-polarization signals tend to be similar, involving a low α value, typical for surface reflection. This scattering mechanism is weighted by a strong intensity and dominates secondary intensities, potentially corresponding to multiple scattering terms, and results in a very low entropy value. This nonstationary behavior was found to have a preponderant influence on the polarimetric properties of resonating field at full resolution. Here, α and H values are significantly lower than those for similar fields that remained unaffected by Bragg resonance. The oblique resonating stripes, shown in the different range-frequency planes in Fig. 5.12, illus-

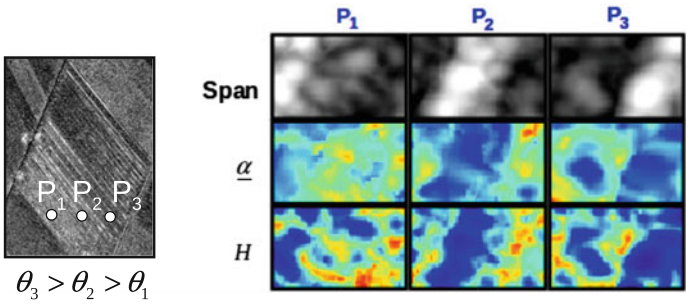


Fig. 5.12 Location of test points P1, P2, and P3 and range-azimuth frequency representation plane (*left*) and representation of polarimetric characteristics in this domain

trate well the dependence of the resonance condition on both range and azimuth frequencies, as shown in Fig. 5.11. It can also be observed that as the incidence angle increases, from $P1$ to $P3$, the oblique resonating stripe slides from low azimuth frequencies to higher ones. This displacement of the resonance locations is due to the dependence of the Bragg condition on the incidence angle and corroborates the analysis of the Bragg resonance as presented in Fig. 5.11. Polarimetric indicators of pixels that do not belong to resonating stripes are unaffected by the Bragg resonance and have values similar to those observed over stationary fields.

5.7.2.2 Detection of Non-stationary Polarimetric TF Behaviors

Each pixel of the SAR scene is associated with a set of R independent target vectors, $\mathbf{k}(\boldsymbol{\omega}_i)$ with $i = 1, \ldots, R$, derived from independent range-azimuth subspectra, i.e., subspectra selected using non-overlapping functions $G(\boldsymbol{\omega}_i)$. Under the classical speckle affected scattering hypothesis, these target vectors follow independent complex Gaussian multivariate distributions, $f(\mathbf{k}(\boldsymbol{\omega}_i)) = \mathscr{N}_{\mathscr{C}}(\mathbf{0}, \boldsymbol{\Sigma}_i)$. The stationary aspect of the scattering behavior of each pixel may then be studied by comparing the second order statistics of $\mathbf{k}(\boldsymbol{\omega}_i)$ for different spectral locations, i.e., by testing the following hypothesis:

$$H : \boldsymbol{\Sigma}_1 = \cdots \boldsymbol{\Sigma}_R = \boldsymbol{\Sigma} \tag{5.78}$$

As it is shown in [42] this hypothesis can be tested easily using sample coherency matrices obtained from n_i independent realizations or looks of each TF target vector, $\mathbf{k}(\boldsymbol{\omega}_i)$:

$$\mathbf{T}_i = \frac{1}{n_i} \sum_{l=1}^{n_i} \mathbf{k}_l(\boldsymbol{\omega}_i) \mathbf{k}_l^H(\boldsymbol{\omega}_i) \quad f(\mathbf{T}_i | \boldsymbol{\Sigma}_i) = \mathscr{W}_{\mathscr{C}}(n_i, \boldsymbol{\Sigma}_i) \tag{5.79}$$

The TF stationary behavior of $\mathbf{k}(\boldsymbol{\omega}_i)$ is evaluated by means of a Maximum Likelihood (ML) ratio, Λ, built from the R independent sample coherency matrices as follows:

$$\Lambda = \frac{\max\, p(\mathbf{T}_1, \ldots, \mathbf{T}_R | H \text{ true})}{\max\, p(\mathbf{T}_1, \ldots, \mathbf{T}_R | H \text{ false})} = \frac{\max\limits_{\boldsymbol{\Sigma}} p(\mathbf{T}_1 | \boldsymbol{\Sigma}, \ldots, \mathbf{T}_R | \boldsymbol{\Sigma})}{\max\limits_{\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_R} p(\mathbf{T}_1 | \boldsymbol{\Sigma}_1, \ldots, \mathbf{T}_R | \boldsymbol{\Sigma}_R)} \tag{5.80}$$

Replacing the likelihoods in (5.80) by their expression and the expectations by their ML estimates, on gets the following simple expression [42]

$$\Lambda = \frac{\prod\limits_{i=1}^{R} \det(\mathbf{T}_i)^{n_i}}{\det(\mathbf{T}_t)^{n_t}} \quad \text{with} \quad \mathbf{T}_t = \frac{1}{n_t} \sum_{i=1}^{R} n_i \mathbf{T}_i \text{ and } n_t = \sum_{i=1}^{R} n_i \tag{5.81}$$

The hypothesis is accepted and the pixel under test is considered to have a stationary polarimetric TF behavior, with an arbitrarily chosen probability of false alarm Pfa, if $\Lambda > c_\beta$, where the relation between the threshold value and the probability of false alarm, $P_{fa}(c_\beta) = \beta$, has been derived in [42]. The ML ratio and nonstationary pixel map shown in Fig. 5.13 indicate that a significant number of pixels have a nonstationary behavior during the duration of the SAR acquisition. Most of the varying scatterers belong to agricultural fields affected by Bragg resonance. Complex targets and diffracting edges, whose scattering characteristics highly depend on the observation position, are discriminated over built-up areas and linear alignment of scatterers.

The ML ratio based detection approach may be further developed to determine nonstationary scattering behavior position in the range-Doppler spectrum by comparing the contributions of each subspectrum image in the global ML ratio information [42]. It can be observed from the localization results displayed in Fig. 5.14 on many fields affected by Bragg resonance that some groups of pixels, belonging to the same

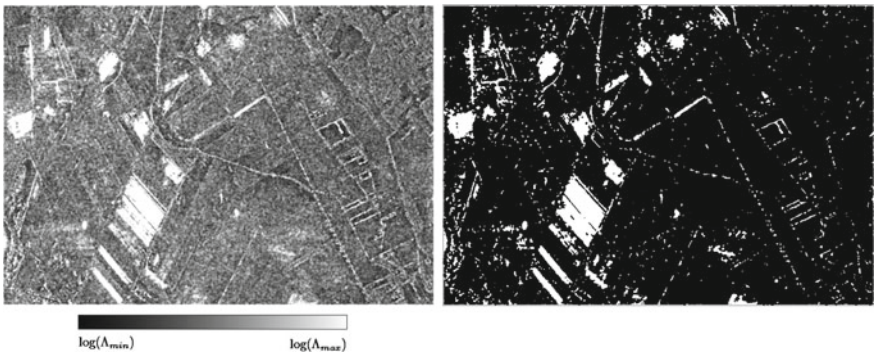


Fig. 5.13 Discrimination of non-stationary scatterers. Image of the ML ratio in log-scale (*left*), non-stationary pixel map (*right*). Non stationary pixels are represented in white

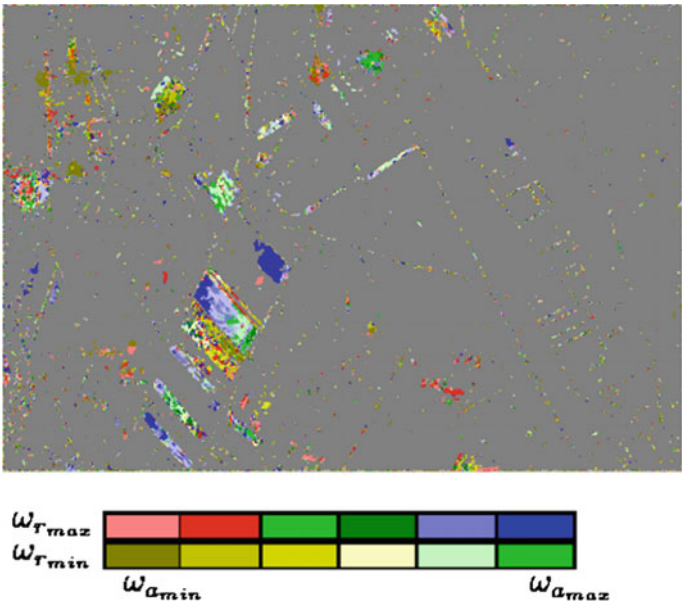


Fig. 5.14 Spectral location of the least stationary component among 12 range-azimuth subspectra for each nonstationary pixel

field, have a maximum anisotropic behavior in different subapertures. This is a consequence of the sliding effects of Bragg resonance on periodic structures, that is described in the next section.

5.7.3 TF Polarimetric Characterization of Complex Scenes

5.7.3.1 Polarimetric Time-Frequency Features

Figure 5.15 shows a color-coded polarimetric SAR image of the city of Dresden acquired by DLR's E-SAR sensor data at L-band. The scene is mainly composed of built-up areas including vegetation spots. A forest and a park can be seen on the left part of the image and a river with smooth banks is located on the right part.

Polarimetric properties of media are generally investigated through a decomposition of second order multivariate polarimetric representations. The resulting parameters provide information on the media geometrical structure and on the underlying scattering mechanisms. Two parameters, obtained from the well known eigenvector-based decomposition introduced in [46] are displayed in Fig. 5.16.

The entropy image shown in Fig. 5.16 reveals that the polarimetric behavior of most of the scene is highly random. Over urban areas, the polarimetric response is

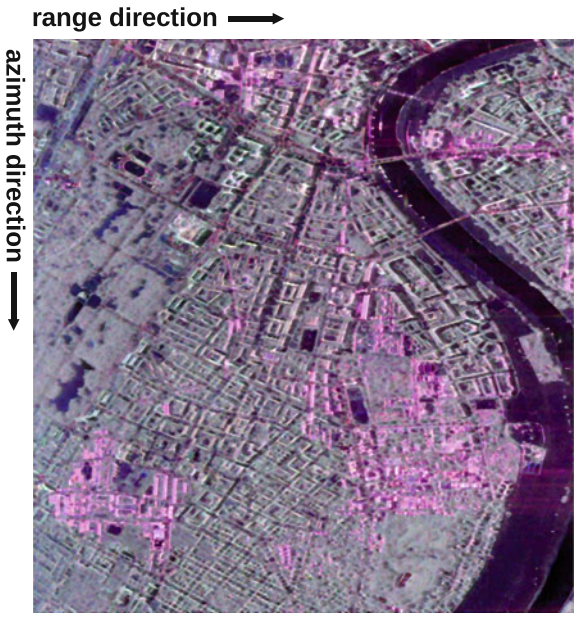


Fig. 5.15 Color coded image of the Dresden test site (Pauli basis)

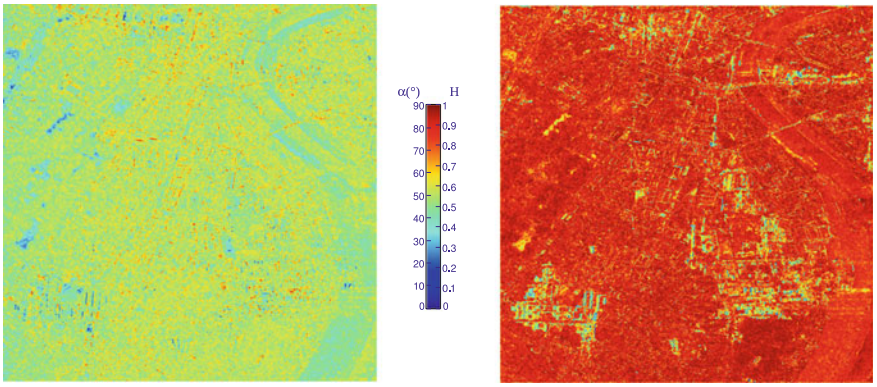


Fig. 5.16 Polarimetric parameter images, $\underline{\alpha}$ (*left*) and H (*right*)

composed of a large number of different polarimetric contributions originating from complex building structures as well as from surrounding vegetation. The resulting high entropy implies that an interpretation of polarimetric indicators may not be relevant. Over buildings aligned with the flight track direction, the entropy has intermediate values and the α parameter reveals the presence of dominant single and double bounce reflexions. Buildings that do not face the radar track are characterized

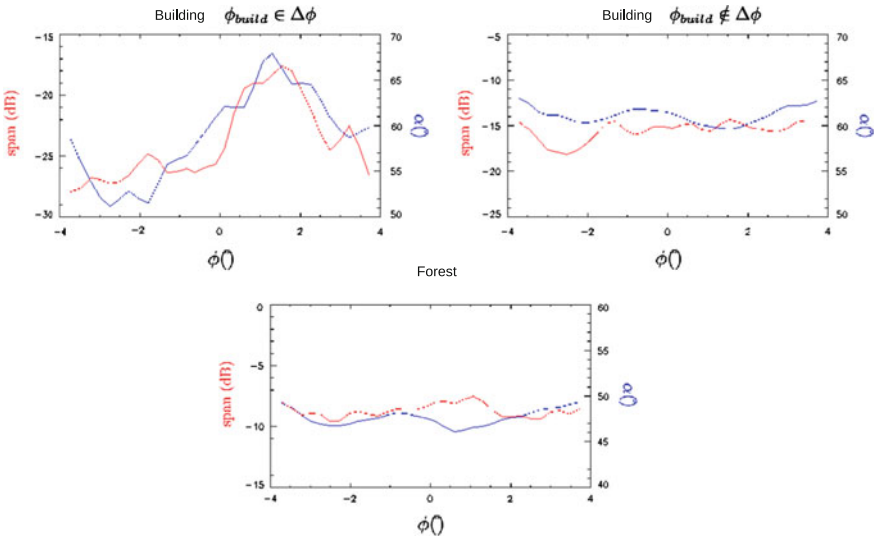


Fig. 5.17 Continuous TF analysis in the azimuth direction (SPAN in *red*, $\underline{\alpha}$ in *blue*) (color figure online)

by a strong cross-polarization component and high entropy and can hardly be discriminated from vegetated areas.

Figure 5.17 presents a continuous TF analysis in the azimuth direction of three different media: a building facing the radar, an oriented building, and a forested area. The SPAN, corresponding to the sum of the intensities in all polarimetric channels, and the polarimetric α angle are computed for the different media at each frequency location and mean values are then estimated over pixels belonging to the object.

A non stationary behavior is clearly visible in Fig. 5.17 with a sudden large variation of both SPAN and α levels with the observation angle in azimuth ϕ. This anisotropic behavior is due to the highly directional patterns of coherent scattering mechanisms which may occur as the radar faces a large artificial structure, such as a building [47]. This particular effect can only be observed if the building orientation with respect to the radar flight track falls within the processed antenna azimuth aperture.

On the contrary, oriented buildings and vegetated areas, like forest patches, show stationary behaviors. The identification of buildings from their TF response thus requires an additional criterion to complement the stationarity information. It is known that man-made objects are likely to have a coherent response, whereas natural media may be considered as random. The discrimination of such responses can be achieved by studying the coherence of the backscattered polarimetric signal in the Time-Frequency domain and requires the use of an adequate TF polarimetric SAR (PolSAR) signal model.

5.7.3.2 PolSAR TF Signal Modeling and Analysis

The proposed TF signal model [44, 48] is given by the following expression, where the spatial coordinates, $\mathbf{l}$, have been omitted:

$$\mathbf{s}(\boldsymbol{\omega}) = \mathbf{t}(\boldsymbol{\omega}) + \mathbf{c}(\boldsymbol{\omega}) \tag{5.82}$$

The signal $\mathbf{s}(\boldsymbol{\omega})$ contains the full coherent polarimetric polarization information and can be associated to a well-known scattering vector [46]:

$$\mathbf{k}(\boldsymbol{\omega}) = \frac{1}{\sqrt{2}} [S_{hh}(\boldsymbol{\omega}) + S_{vv}(\boldsymbol{\omega}), S_{hh}(\boldsymbol{\omega}) + S_{vv}(\boldsymbol{\omega}), 2S_{hv}(\boldsymbol{\omega})]^T \tag{5.83}$$

where $S_{pq}(\boldsymbol{\omega})$ represents an element of the (2×2) scattering matrix $\mathbf{S}$ sampled at the frequency coordinates $\boldsymbol{\omega}$.

The signal described in (5.82) is composed of two contributions:

- The term $\mathbf{t}(\boldsymbol{\omega})$ is highly coherent and can be associated to a deterministic or almost deterministic target response. Depending on the structure of the observed object, the response can remain constant during the SAR acquisition, or can be non-stationary if the backscattering behavior is sensitive to the azimuth angle of observation or illumination frequency.
- The second term, $\mathbf{c}(\boldsymbol{\omega})$, represents the response of distributed environments. It is uncorrelated, but may follow a non-stationary behavior in particular cases, e.g., vegetated terrains with a strong topography, very dense environments whose response results from the sum of a large number of uncorrelated contributions.

This composite model may be tested using $\mathbf{s}(\boldsymbol{\omega})$ second order statistics:

- The coherence of $\mathbf{s}(\boldsymbol{\omega})$ can be used to determine the dominant component within the pixel under consideration. A high value indicates that $\mathbf{t}(\boldsymbol{\omega})$ is the most important term in (5.82), and a low one corresponds to scattering from an incoherent, distributed, medium.
- The stability of the dominant component can then be tested by studying the stationarity of the variance of $\mathbf{s}(\boldsymbol{\omega})$

Due to the signal high dimensionality, usual scalar tools are not well adapted to the study of second-order TF polarimetric statistics. A polarimetric TF target vector is built by gathering the PolSAR information sampled at R spectral coordinates $\boldsymbol{\omega}_i$, $i = 1, \ldots, R$.

$$\mathbf{k_{TF}} = \left[\mathbf{k}^T(\boldsymbol{\omega}_1), \ldots, \mathbf{k}^T(\boldsymbol{\omega}_R)\right]^T \tag{5.84}$$

The sampling coordinates, $\boldsymbol{\omega}_i$, and the frequency domain resolution of the analyzing function g are chosen so that the R sub-spectra do not overlap and span the whole full resolution spectrum [43]. A polarimetric TF sample covariance matrix, $\mathbf{T_{TF-Pol}}$, is then computed as follows

$$\mathbf{T_{TF-Pol}} = \left\langle \mathbf{k_{TF}} \mathbf{k_{TF}}^{\dagger} \right\rangle = \begin{bmatrix} \mathbf{T_{11}} & \cdots & \mathbf{T_{1R}} \\ \vdots & \ddots & \vdots \\ \mathbf{T_{R1}} & \cdots & \mathbf{T_{RR}} \end{bmatrix}$$
$$\text{where } \mathbf{T_{ij}} = \left\langle \mathbf{k}(\boldsymbol{\omega}_i)\mathbf{k}(\boldsymbol{\omega}_j)^{\dagger} \right\rangle \tag{5.85}$$

Stationarity is assessed by testing the fluctuations of the variance of the signal at the different spectral locations [42, 43] as shown in (5.81).

Figure 5.18 presents a log-image of the Λ parameter on the Dresden test site, obtained with 4 spectral coordinates in the azimuth direction over the Dresden test site.

The Λ parameter reached high values over natural areas indicating a stationary spectral behavior. Over buildings, Λ decreases, pointing out the invalidity of the stationary hypothesis over such objects. Highly anisotropic pixels, such as those corresponding to the wall-ground dihedral reflection or specular reflection from oriented roofs are clearly identified in Fig. 5.18 due to their very low stationary aspect.

In [49], the eigenvalues of a single-polarization covariance matrix have been used to derive a coherency indicator. These eigenvalues carry information on the

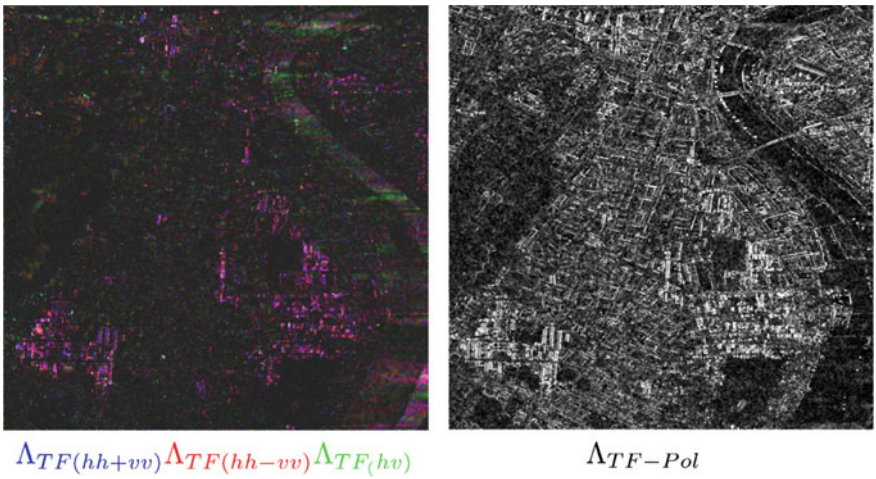


Fig. 5.18 Non stationary TF behavior indicator, Λ_{TF}, computed separately for different polarimetric channels (*left*) and simultaneously using the whole polarimetric information (*right*)

correlation structure, but are also sensitive to potential PolSAR fluctuations due to non-stationarity. A solution has been proposed in order to overcome this limitation and to jointly use all the polarimetric channels [48]. Under the hypothesis of uncorrelated spectral responses, the off-diagonal terms of the TF covariance matrix verify:

$$H_\theta : \boldsymbol{\Sigma}_{ij} = \mathbf{0} \quad \forall i \neq j \tag{5.86}$$

The corresponding ML ratio is given by:

$$\boldsymbol{\Theta} = \frac{\max_{\boldsymbol{\Sigma}_{ii}} L(\boldsymbol{\Sigma}_{11}, \ldots, \boldsymbol{\Sigma}_{RR})}{\max_{\boldsymbol{\Sigma}_{TF}} L(\boldsymbol{\Sigma}_{TF})} = \frac{|\mathbf{T}_{TF}|^{n_i}}{\prod_{i=1}^{R} |\mathbf{T}_{ii}|^{n_i}} \tag{5.87}$$

This ML ratio expression can be rewritten as $\boldsymbol{\Theta} = \left|\tilde{\mathbf{T}}_{TF}\right|^{n_i}$ with

$$\tilde{\mathbf{T}}_{\mathbf{TF-Pol}} = \begin{bmatrix} \mathbf{I} & \boldsymbol{\Gamma}_{12} & \cdots & \boldsymbol{\Gamma}_{1R} \\ \boldsymbol{\Gamma}_{12}^{\dagger} & \mathbf{I} & & \vdots \\ \vdots & & \ddots & \vdots \\ \boldsymbol{\Gamma}_{1R}^{\dagger} & \cdots & \cdots & \mathbf{I} \end{bmatrix} \text{ where } \boldsymbol{\Gamma}_{ij} = \mathbf{T_{ii}}^{-1/2} \mathbf{T_{ij}} \mathbf{T_{jj}}^{-1/2} \tag{5.88}$$

The normalized covariance matrix, $\tilde{\mathbf{T}}_{\mathbf{TF-Pol}}$ results from the whitening of the TF polarimetric covariance matrix by the separate polarimetric information at each frequency location. This representation is then insensitive to spectral polarimetric intensity variations and is characterized by its off-diagonal matrices $\boldsymbol{\Gamma}_{ij}$ which can be viewed as an extension of the scalar normalized correlation coefficient to the polarimetric case. The ML ratio in (5.87) is a function of the eigenvalues of $\tilde{\mathbf{T}}_{\mathbf{TF-Pol}}$, which reflect the correlation structure: flat for decorrelated responses ($\tilde{\mathbf{T}}_{\mathbf{TF-Pol}} \rightarrow \mathbf{I_d}$), heterogeneous for correlated ones. Taking into account $\tilde{\mathbf{T}}_{\mathbf{TF-Pol}}$ peculiar form, a correlation indicator, named TF-Pol coherence, can be defined as [48]:

$$\rho_{TF-Pol} = 1 - \left|\tilde{\mathbf{T}}_{\mathbf{TF-Pol}}\right|^{\frac{1}{3R}} \tag{5.89}$$

Figure 5.19 presents an image of ρ_{TF-Pol} over the Dresden test site, computed from 4 spectral locations in the azimuth direction.

As expected, the TF-Pol coherence is high over buildings due to the presence of strong coherent reflectors. It can be also noticed that buildings are identified independently of their orientation.

5.7.3.3 Scene Analysis Using TF Stationarity and Coherence

The stationarity and coherence indicators derived above can be merged to classify the scene. Both ρ_{TF-Pol} and Λ parameters are thresholded and combined into four

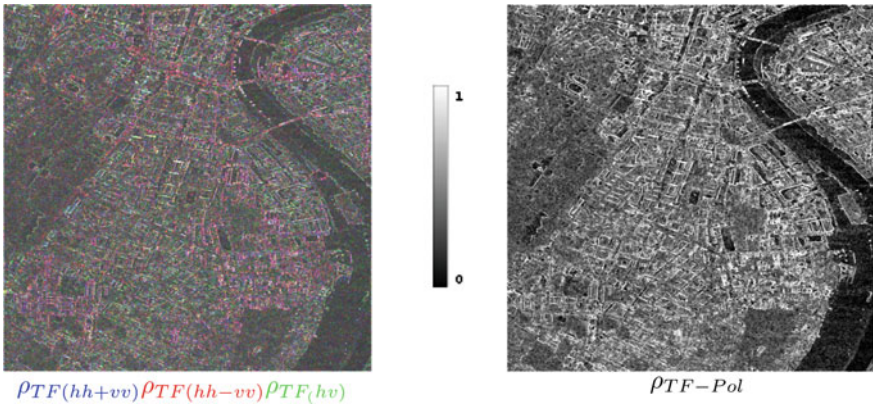


Fig. 5.19 TF coherence indicator, ρ_{TF}, computed separately for different polarimetric channels (*left*) simultaneously using the whole polarimetric information (*right*)

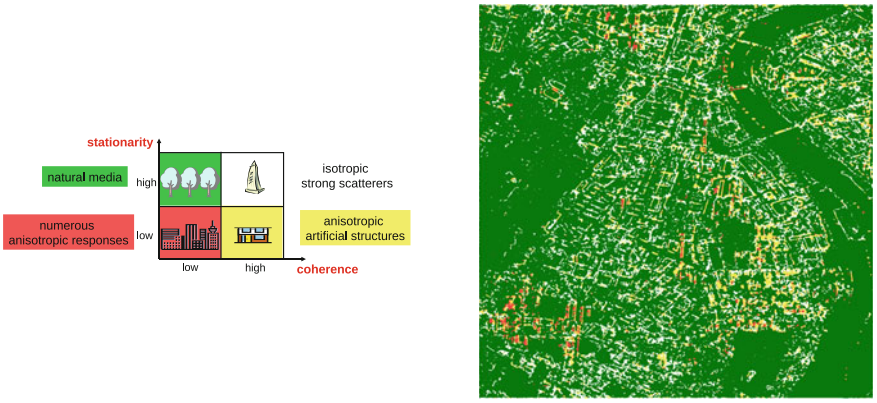


Fig. 5.20 TF polarimetric classification. Classification scheme (*left*) and results obtained over Dresden (*right*)

classes. The application of the fusion strategy over the Dresden site is shown in Fig. 5.20.

The resulting map permits a good estimation of building locations. A physical interpretation can be given for each of the four classes:

- Coherent and stationary pixels (white class): The $\mathbf{t}$ term in (5.82) is dominant and constant during the SAR acquisition. This kind of behavior corresponds to strong scatterers with an isotropic response, like oriented buildings, lamp-posts, ….

- Coherent and non stationary pixels (yellow class): The **t** contribution dominates but varies during the measure, causing fluctuation of the signal with the azimuth angle of observation. This anisotropic effect is characteristic of buildings facing the radar track, whose response is affected by a strong and highly directive pattern, mainly due to double bounce reflections or specular single bounce reflection over roofs tilted toward the radar.
- Incoherent and stationarity pixels (green class): The uncorrelated component **c** dominates and has stable second order statistics. This class corresponds to natural environments (forests, fields, grass areas, …) of distributed artificial media such as roads, roof tops, terraces, …
- Incoherent and non-stationary pixels (red class): This class indicates the presence of complex scattering contributions, which change during the SAR integration, and sum-up in an incoherent way, like in layover areas.

As it was shown in Fig. 5.16, the full resolution PolSAR information can't really be used to analyze the scene geophysical properties due to a very high entropy inherent to the study of dense environments. The proposed PolSAR TF analysis technique can also be used to improve in a significant way the interpretation of polarimetric indicators. The most coherent TF scattering mechanisms is described by the first eigenvector of $\tilde{\mathbf{T}}_{\mathbf{TF-Pol}}$, which can be transformed back to the H-V polarimetric basis using a matrix $\mathbf{P}$, satisfying $\tilde{\mathbf{T}}_{\mathbf{TF-Pol}} = \mathbf{P}\mathbf{T}_{\mathbf{TF-Pol}}\mathbf{P}^{\dagger}$. From this eigenvector, one can extract an α_{TF} parameter which shows a much more contrasted and relevant information than the original full resolution parameter α. Figure 5.21 shows different images of a building of the scene. A comparison between the T-F classification results and the optical image reveals that the double bounce reflection is considered as both a non-stationary and coherent scattering mechanism, whereas the roof layover is seen as non-stationary and uncorrelated, due to the superposition of the roof and ground contributions. A thresholding of α_{TF} with respect to $\pi/4$, permits us to easily separate these two different mechanisms and could be used to get a rough estimate of the building height. Such information might be useful to interferometric phase unwrapping algorithms which generally face ambiguity issues over urban areas.

5.7.3.4 Target Detection Using Polarimetric SAR Images Acquired by Spaceborne Sensors

The proposed polarimetric TF characterization technique may be used to discriminate targets from their background by detecting and extracting coherent components in complex random SAR responses acquired by spaceborne sensors. It is known that SAR devices operating from space generally:

- have a very narrow antenna beam in the azimuth direction in order to maintain the pulse emission frequency to a low value
- emit signals over a restricted spectral bandwidth

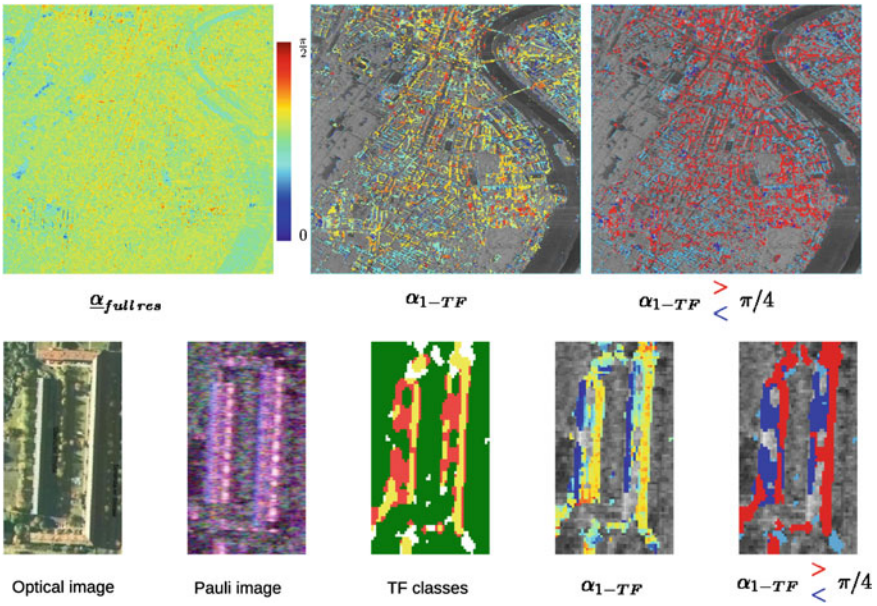


Fig. 5.21 Application of the classification scheme to building characterization: over the whole image (*top*), over the an isolated building (*bottom*)

Due to the resulting very small domain of variation of $\boldsymbol{\omega}$, such SAR data sets may not be used to detect scatterers having a non-stationary TF scattering behavior. Nevertheless, several studies [42, 44, 50–53] have shown that various kinds of artificial environments, such as built-up structures, vehicles, lamp posts, barriers, and so on, had a partially coherent SAR response due to the presence of deterministic components, generally associated with specular single- or multiple-scattering terms. This property may be exploited in order to detect targets using the polarimetric TF coherence indicator presented in (5.89), computed in both azimuth and range directions with an arbitrary number of spectral locations. We present the case for ship detection using a Constant False Alarm Rate (CFAR) detection approaches.

A case study of a complex sea area containing ships, imaged by the RADARSAT-2 sensor operating at C band in Fine-Quad (FQ) polarization mode, is investigated here and is illustrated in Fig. 5.22a, which shows a color-coded full-resolution polarimetric SAR image of the sea harbor area of Vancouver city, Canada, onto which some target (T) and ghost (G) focused echoes are indicated. This image indicates that in coastal areas and harbors, man-made structures over sea and land may cause significant range artifact features, like the ghost indicated as G1 caused by the ship labeled T1 or ghosts G2a and G2b resulting from a specific scattering structure noted T2 which is located in the urban area. Moreover, some potentially metallic structures and ships (such as ships T3 and T5) may have highly energetic responses whose side-lobe intensity lies well above the level of reflectivity of the sea. Such artifacts and scattering patterns

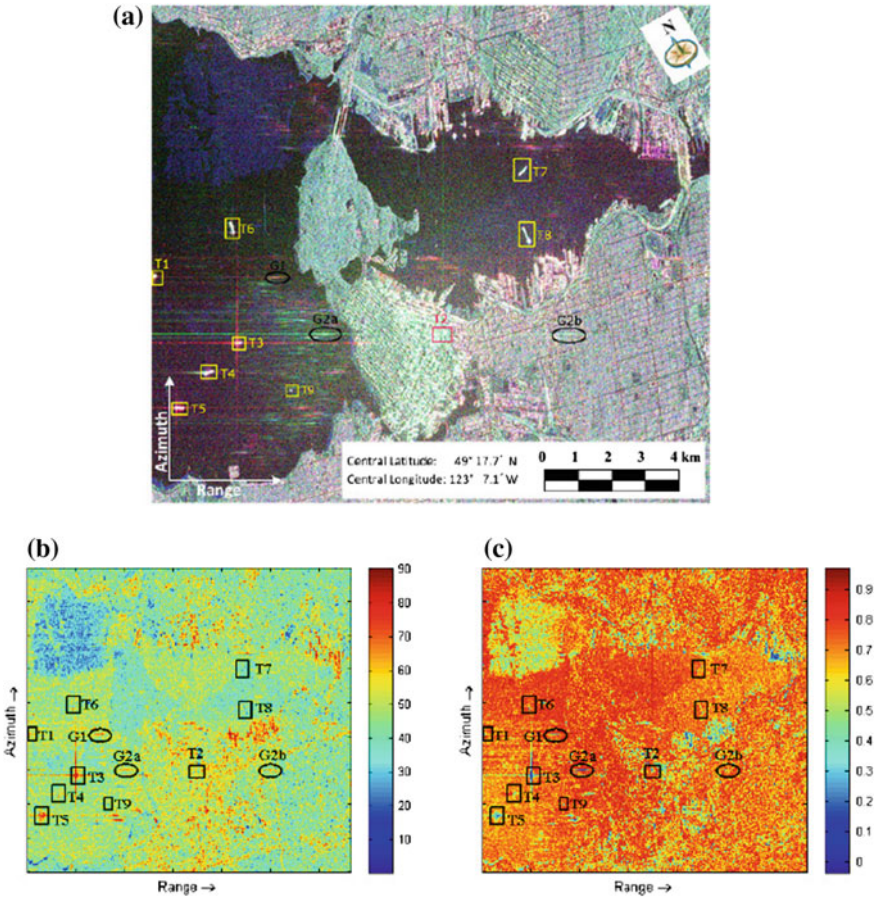


Fig. 5.22 Full resolution polarimetric features of a RADARSAT-2 image acquired over the Vancouver harbor area (Canada). **a** Polarimetric color-coded image (HH + VV, HV + VH, HH − VV). **b** polarimetric entropy H. **c** polarimetric α angle. Some ships Ti and ghosts Gi are indicated by rectangles and ellipses, respectively

are susceptible to generate numerous false alarms if classical detection approaches based on contrasts are applied on such a complex data set. The entropy image shown in Fig. 5.22b reveals that for this scene the polarimetric behavior of most of the sea background is highly random, i.e., its polarimetric covariance matrix tends to be isotropic or white, and polarimetrically adaptive detection schemes may have limited performance. Some ships, such as T1, T4, T6, T7, T8, and T9, also have a high degree of polarimetric randomness, due to the mixing of different polarimetric contributions originating from their complex structures as well as from the surrounding sea area and to potentially superimposed artifacts. Due to their high entropy and intermediate α value, they can hardly be discriminated from the surrounding sea. Other ships,

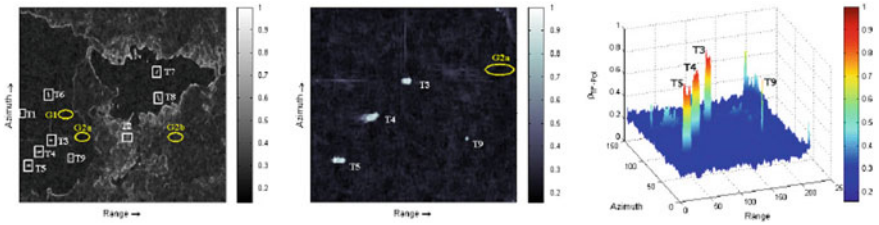


Fig. 5.23 Polarimetric TF coherence γ_{TF-Pol} - evaluated over the Vancouver harbor data set in both range and azimuth directions simultaneously. Whole image (*left*) Closeup view (*center*) 3-D closeup plot (*right*)

such as T3 and T5, with a heading direction perpendicular to the azimuth direction, show a low entropy value and their α parameter reveals the presence of dominant double bounce reflections. This almost deterministic polarimetric behavior, due to a scattering level much higher than the one of their environment, is well adapted to contrast-based detection. Nevertheless, the resulting widespread side-lobes may cause false alarms. One may remark that the ship labeled T1 and its ghost G1 have very similar, very deterministic polarimetric features and hence cannot be discriminated using polarimetric diversity only.

The polarimetric TF coherence analysis is applied to the Vancouver harbor data set presented in Fig. 5.22 for a spectral sampling performed in both azimuth and range directions. The corresponding TF coherence map, obtained from sub-spectra sampled at two frequency locations in each direction, each spectrum occupying 25% of the total available 2D domain is given in Fig. 5.23. Compared with the full-resolution polarimetric features depicted in Fig. 5.22, the γ_{TF} displayed in Fig. 5.23 show the power of discrimination of the proposed TF analysis technique for ship discrimination in severe backgrounds, especially for the ships T1, T4, T6, T7, and T8 which are mixed with different polarimetric contributions originating from their complex structures as well as from the surrounding sea area. Distributed environments like the sea and continental ground have small TF coherence values, whereas ships are recognized as highly coherent scatterers. Moreover, these results indicate that this approach can effectively mitigate artifacts related to SAR ambiguity, as widespread ghosts with high entropy are filtered out due to their incoherent behaviors and point-like ones, such as ghosts G1 and G2a, shown in Fig. 5.22, which could have easily been misinterpreted as ships by conventional discrimination techniques, have been cancelled too. One may note that this technique can reduce cross-shaped side-lobes generated by the ships T3 and T5, which could have been also misclassified as targets. In addition, as shown in Fig. 5.23, besides the larger ships (T3–T8), the smaller ship (T9) can be detected and localized, despite the fact that TF analysis techniques may degrade the spatial resolution. Therefore, the proposed indicator permits to detect ships, even though the sea environment is complex due to ghost echoes and highly random polarimetric behavior of the scene.

5.8 Conclusion

In this chapter we have introduced the fundamental statistical modelling of polarimetric synthetic aperture radar image data at the pixel-variate level and followed some of its mathematical consequences through parameter estimation and Bayesian classification. Furthermore, we explored a deeper level of sub-aperture time-frequency analysis to detect signals with time or angular varying properties. The two approaches are complementary and both involve strong mathematical modelling that helps us interpret SAR and PolSAR images.

References

1. Ishimaru, A.: Wave Propagation and Scattering in Random Media. Academic Press Inc, New York (1978)
2. Goodman, J.W.: Some fundamental properties of speckle. J. Opt. Soc. Am. **66**(11), 1145–1150 (1976)
3. Whalen, A.D.: Detection of Signals in Noise. Academic Press Inc, New York (1971)
4. Oliver, C., Quegan, S.: Understanding Synthetic Aperture Radar Images. SciTech Publishing Inc, Raleigh (2004)
5. Broadbent, S.R., Kendall, D.G.: The random walk of trichostrongylus retortaeformis. Biometrics **9**, 460–466 (1953)
6. Yasuda, N.: The random walk model of human migration. Theor. Popul. Biol. **7**, 156–157 (1975)
7. Jakeman, E., Pusey, P.N.: Significance of k distributions in scattering experiments. Phys. Rev. Lett. **40**, 546–550 (1978)
8. Oliver, C.J.: A model for non-rayleigh scattering statistics. Opt. Acta. **31**(6), 701–722 (1984)
9. Jakeman, E., Pusey, P.N.: A model for non-rayleigh sea echo. IEEE Trans. Antennas Propagat. **AP–31**(4), 490–498 (1975)
10. Jakeman, E.: Speckle statistics with a small number of scatterers. Opt. Eng. **23**(4), 453–461 (1984)
11. Eltoft, T., Høgda, K.A.: Non-gaussian signal statistics in ocean sar imagery. IEEE Trans. Geosci. Remote Sens. **36**(2), 562–575 (1998)
12. Andrews, A.F., Mallow, C.L.: Scale mixtures of normal distributions. J. Roy. Stat. Soc. B **36**(1), 99–102 (1974)
13. Lee, J.-S., Pottier, E.: Polarimetric Radar Imaging, From Basics to Applications. Taylor & Francis Group (2009)
14. Cloude, S.R.: Polarisation: Applications in Remote Sensing. Oxford University Press, Oxford (2010)
15. Boerner, W.-M.: Basic concepts in radar polarimetry. Technical report, UIC-ECE Communications, Sensing and Navigation Laboratory, Chicago, IL/USA (2010)
16. Gini, F., Greco, M.: Covariance matrix estimation for CFAR detection in correlated heavy tailed clutter. Elsevier Signal Proc. **82**(12), 1847–1859 (2002)
17. Tao, D., Anfinsen, S.N., Brekke, C.: A comparative study of sea clutter covariance matrix estimators. IEEE Geosci. Remote Sens. Lett. **11**(5), 1010–1014 (2014)
18. Tyler, D.E.: A distribution-free M-estimator of multivariate scatter. Ann. Statist. **15**(1), 234–251 (1987)
19. Conte, E., De Maio, A., Ricci, G.: Recursive estimation of the covariance matrix of a compound-Gaussian process and its application to adaptive CFAR detection. IEEE Trans. Signal Process. **50**(8), 1908–1915 (2002)

20. Pascal, F., Forster, P., Ovarlez, J.-P., Larzabal, P.: Performance analysis of covariance matrix estimates in impulsive noise. IEEE Trans. Geosci. Remote Sens. **56**(6), 2206–2217 (2008)
21. Pascal, F., Chitour, Y., Forster, P., Larzabal, P.: Covariance structure maximum-likelihood estimates in compound gaussian noise: existence and algorithm analysis. IEEE Trans. Signal Process. **56**(1), 34–48 (2008)
22. Vasile, G., Ovarlez, J.-P., Pascal, F., Tison, C.: Coherency matrix estimation of heterogeneous clutter in high-resolution polarimetric SAR images. IEEE Trans. Geosci. Remote Sens. **48**(4), 1809–1826 (2010)
23. Mathai, A.M.: Jacobians of Matrix Transformations and Functions of Matrix Arguments. World Scientific, New York (1997)
24. Nicolas, J.-M.: Introduction aux statistique de deuxième espèce: Application des logs-moments et des logs-cumulants à l'analyse des lois d'images radar. Traitement du Signal **19**(3), 139–167 (2002). In French, English translation in [54]
25. Anfinsen, S.N., Eltoft, T.: Application of the matrix-variate Mellin transform to analysis of polarimetric radar images. IEEE Trans. Geosci. Remote Sens. **49**(6), 2281–2295 (2011)
26. Oliver, C., Quegan, S.: Understanding Synthetic Aperture Radar Images, 2nd edn. SciTech Publishing, Raleigh (2004)
27. Anfinsen, S.N., Doulgeris, A.P., Eltoft, T.: Estimation of the equivalent number of looks in polarimetric synthetic aperture radar imagery. IEEE Trans. Geosci. Remote Sens. **47**(11), 3795–3809 (2009)
28. Maiwald, D., Kraus, D.: Calculation of moments of complex Wishart and complex inverse Wishart distributed matrices. IEE Proc. Radar, Sonar, Navigation **147**(4), 162–168 (2000)
29. Foucher, S., Boucher, J.-M., Benie, G.B.: Maximum likelihood estimation of the number of looks in SAR images. In: International Conference on Microwaves, Radar and Wireless Communications, vol. 2, pp. 657–660. Wroclaw, Poland (2000)
30. Tao, L., Cui, H.-G., Xi, Z.-M., Gao, J.: Texture-invariant estimation of equivalent number of looks based on trace moments in polarimetric radar imagery. IEEE Geosci. Remote Sens. Lett. **11**(6), 1129–1133 (2014)
31. Hu, D., Anfinsen, S.N., Qiu, X., Doulgeris, A.P., Leim B.: Unsupervised mixture-eliminating estimation of equivalent number of looks for PolSAR data. IEEE Trans. Geosci. Remote Sens. (2016)
32. Stuart, A., Ord, J.K.: Kendall's Advanced Theory of Statistics: Distribution Theory, vol. 1, 6th edn. Edward Arnold, London (1994)
33. Anfinsen, S.N., Doulgeris, A.P., Eltoft, T.: Goodness-of-fit tests for multilook polarimetric radar data based on the Mellin transform. IEEE Trans. Geosci. Remote Sens. **49**(8), 2764–2781 (2011)
34. http://mdacorporation.com/geospatial/international/satellites/radarsat-2/sample-data. MDA corporation sample data web-site
35. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Stat. Soc. B **39**(1), 1–38 (1977)
36. McLachlan, G.J., Peel, D.: Finite Mixture Models. Wiley, New York (2000)
37. Doulgeris, A.P.: An automatic U-distribution and Markov random field segmentation algorithm for PolSAR images. IEEE Trans. Geosci. Remote Sens. **53**(4), 1819–1827 (2015)
38. Leducq, P., Ferro-Famil, L., Pottier, E.: Matching-pursuit-based analysis of moving objects in polarimetric sar images. IEEE Geosci. Remote Sens. Lett. **5**(2), 123–127 (2008)
39. Bamler, R., Hartl, P.: Synthetic aperture radar interferometry. Inverse Probl. **14**(4) (1998)
40. Gatelli, F., Guamieri, A.M., Parizzi, F., Pasquali, P., Prati, C., Rocca, F.: The wavenumber shift in sar interferometry. IEEE Trans. Geosci. Remote Sens. **32**(4), 855–865 (1994)
41. Flandrin, P.: Time-Frequency/Time-Scale Analysis. Wavelet Anlysis Appl. Academic Press, London (1999)
42. Ferro-Famil, L., Reigber, A., Pottier, E., Boerner, W.M.: Scene characterization using subaperture polarimetric sar data. IEEE Trans. Geosci. Remote Sens. **41**(10), 2264–2276 (2003)
43. Ferro-Famil, L., Reigber, A., Pottier, E.: Nonstationary natural media analysis from polarimetric sar data using a two-dimensional time-frequency decomposition approach. Can. J. Remote Sens. **31**(1), 20–29 (2005)

44. Hu, C., Ferro-Famil, L., Kuang, G.: Ship discrimination using polarimetric SAR data and coherent time-frequency analysis. Remote Sens. **5**(12), 6899–6920 (2013)
45. Yueh, A., Shin, R.T., Kong, J.A.: Scattering from randomly perturbed periodic and quasiperiodic surfaces. Prog. Electromagn. Res. **1**, 297–358 (1988)
46. Cloude, S.R., Pottier, E.: An entropy based classification scheme for land applications of polarimetric sar. IEEE Trans. Geosci. Remote Sens. **35**(1), 68–78 (1997)
47. Franceschetti, G., Iodice, A., Riccio, D., Ruello, G.: Sar raw signal simulation for urban structures. IEEE Trans. Geosci. Remote Sens. **41**(9), 1986–1995 (2003)
48. Ferro-Famil, L., Pottier, E.: Urban area remote sensing from l- band polsar data using time-frequency techniques. In: Proceedings Urban Conference (2007)
49. Schneider, R.Z., Papathanassiou, K.P., Hajnsek, I., Moreira, A.: Polarimetric and interferometric characterization of coherent scatterers in urban areas. IEEE Trans. Geosci. Remote Sens. **44**(4), 971–984 (2006)
50. Sauer, S., Ferro-Famil, L., Reigber, A., Pottier, E.: Polarimetric dual-baseline insar building height estimation at l-band. IEEE Geosci. Remote Sens. Lett. **6**(3), 408–412 (2009)
51. Huang, Y., Ferro-Famil, L., Reigber, A.: Object imaging using SAR tomography and polarimetric spectral estimators. IEEE Trans. Geosci. Remote Sens. **50**(6), 2213–2225 (2012)
52. Navarro-Sanchez, V.D., Lopez-Sanchez, J.M., Ferro-Famil, L.: Polarimetric approaches for persistent scatterers interferometry. IEEE Trans. Geosci. Remote Sens. **52**(3), 1667–1676 (2014)
53. Ferro-Famil, L., Huang, Y., Pottier, E.: Principles and applications of polarimetric SAR tomography for the characterization of complex environments. In: Sanso, F. (ed.) International Association of Geodesy Symposia, vol. 142(1–13), Springer, Berlin (2015)
54. Anfinsen, S.N.: Statistical Analysis of Multilook Polarimetric Radar Images with the Mellin Transform. Ph.D. thesis, University of Tromsø, Tromsø, Norway (2010)

Chapter 6
Remote Sensing Data Fusion: Guided Filter-Based Hyperspectral Pansharpening and Graph-Based Feature-Level Fusion

Wenzhi Liao, Jocelyn Chanussot and Wilfried Philips

Abstract Recent advances in remote sensing technology have led to an increased availability of a multitude of satellite and airborne data sources, with increasing resolution. The term resolution here includes spatial and spectral resolutions. Additionally, at lower altitudes, airplanes and Unmanned Aerial Vehicles (UAVs) can deliver very high-resolution data from targeted locations. Remote sensing acquisitions employ both passive (optical and thermal range, multispectral, and hyperspectral) and active devices such as Synthetic Aperture Radar (SAR) and Light Detection and Ranging (LiDAR). Diverse information of the Earth's surface can be obtained from these multiple imaging sources. Optical and SAR characterize the surface of the ground, LiDAR provides the elevation, while multispectral and hyperspectral sensors reveal the material composition. These multisource remote sensing images, once combined/fused together, provide a more comprehensive interpretation of land cover/use (urban and climatic changes), natural disasters (floods, hurricanes, and earthquakes), and potential exploitation (oil fields and minerals). However, automatic interpretation of remote sensing data remains challenging. Two fundamental problems in data fusion of multisource remote sensing images are (1) differences in resolution hamper the ability to fastly interpret multisource remote sensing images and (2) there is no clear methodology yet on combining the diverse information of different data sources. In this chapter, we will introduce our recent solutions for these two problems, with an introduction on signal-level fusion (hyperspectral image pansharpening) first, followed by feature-level fusion (graph-based fusion model for multisource data classification).

W. Liao (✉) · W. Philips
Department of Telecommunications and Information Processing, Ghent University,
Sint-Pietersnieuwstrat 41, 9000 Ghent, Belgium
e-mail: wenzhi.liao@telin.ugent.be

W. Philips
e-mail: philips@telin.ugent.be

J. Chanussot
GIPSA-Lab, Grenoble Institute of Technology, 38031 Grenoble, France
e-mail: jocelyn.chanussot@gipsa-lab.grenoble-inp.fr

G. Moser and J. Zerubia (eds.), *Mathematical Models for Remote Sensing Image Processing*, Signals and Communication Technology,
https://doi.org/10.1007/978-3-319-66330-2_6

6.1 Introduction

Recent advances in the sensor technology of remote sensing have led to an increased availability of acquiring multisource data from the same area. It is obvious that no single data source can always suffice for reliable Earth observation. Multisource data fusion is a technology to exploit the integration (or fusion) of data from multiple sources, in order to improve the decision making that is difficult to achieve by a single data source alone. Fusion of information with different physical characteristics from multiple sources enables improved analysis and interpretation of remote sensing scenes, and has been also applied more widely to other disciplines, such as computer vision, medical imaging, sensor networks, robotics, intelligent system design, etc. Multisource data fusion is also known as multimodal data fusion, includes multisensor data fusion, and is a subset of information fusion.

The current and upcoming Earth observation satellite missions, e.g., ESA Sentinels, NASA A-Train satellite constellation and Jilin constellation of China, allow us to acquire massive remote sensing images of different spatial, spectral, temporal resolutions. In addition, unmanned aerial vehicles (UAVs) can deliver extremely high-resolution data from targeted locations. This huge amount of remote sensing data exacerbate the need to develop techniques for multisource data fusion (the phenomenon that is also referred to as "Big Data" in various fields of science). Generally, multisource data fusion can enhance data authenticity and enable improved detection, confidence, and reliability, as well as reduction in data ambiguity. Meanwhile, fusion of multisource data improves data availability, extending their spatial and temporal coverages.

A general introduction of multisensor data fusion and its applications was provided by Hall and Llinas [1]. A review paper on multiple sensors data fusion techniques [2], explained the concepts, methods, and applications of image fusion as a contribution to multisensor integration-oriented data processing. Since then, image fusion has received increasing attention. Zhang [3] discussed optical panchromatic and multispectral data fusing methods, and reviewed current techniques of multisource remote sensing data fusion, and discusses their future trends and challenges through the concept of hierarchical classification, followed by an overview [4] on data fusion techniques for urban remote sensing. Dalla Mura et al. [5] summarized the Data Fusion Contest of the IEEE Geoscience and Remote Sensing Society since 2006, gave an overview of the current trends, opportunities, and challenges related to the exploitation of multimodal data for Earth observation. Lahat et al. [6] provided general ideas, perspectives, and guidelines as to how to approach data fusion. Gomez–Chova et al. [7] overviewed the current methodologies to integrate multiple and heterogeneous image sources (e.g., multispectral, hyperspectral, radar, multitemporal, and multiangular images) for data classification.

The information from multiple sources can be fused at different levels of representation, depending on the correlations among the sensors, as well as the needs of the system. Generally, multisource data fusion takes place at four different levels of representation [8]: signal-level, pixel-level, feature-level, and symbol- (or decision-) level. Signal-level fusion aims at combining signals from multiple sources to pro-

vide a signal that is typically the same type as the original source but with improved quality. Pixel-level fusion is used to increase the information registered in each pixel of an image, combining multiple images for example. Feature-level fusion employs various features extracted from multisource data to perform fusion. Decision-level fusion typically refers to the combination of decisions from each data source to produce a final decision.

This chapter will first focus on signal-level multisource data fusion: i.e., combining multisource of raw data to produce a new enhanced raw data, which is more informative and synthetic than the original sources, in Sect. 6.2. Section 6.3 will investigate graph-based feature-level fusion models to integrate multisource data for improved performances on remote sensing image classification.

6.2 Hyperspectral Image Pansharpening

The main advantage of hyperspectral (HS) image with respect to multispectral/RGB ones is the more accurate spectral information they provide, which clearly benefits many applications such as unmixing, change detection, object recognition, scene interpretation, and classification (see also Chap. 2). However, imaging systems are designed to balance two competing constraints, namely the spatial resolution and the signal-to-noise ratio (SNR). Hyperspectral systems have reduced bandwidths which require a coarser instantaneous field of view (IFOV) in order to collect enough photons to maintain an acceptable SNR. Panchromatic/RGB color sensors usually have broader spectral bandwidths, which allow for finer spatial resolution by increasing the SNR over a broad spectral band. Thus, to increase either spectral or spatial resolution, one of the two must be sacrificed. As a special branch of multisource data fusion, pansharpening is an image processing technique that might allow analysts to circumvent this trade off, as well as permit preservation of fine resolution and spectral integrity. Hyperspectral image pansharpening aims at improving the spatial quality of a low spatial resolution (LR) MS/HS image by fusing it with a high-resolution (HR) panchromatic/RGB image.

Many pansharpening methods have been developed over the last two decades. In [9], these methods were divided into four categories: component substitution (CS), multiresolution analysis (MRA), Bayesian, and variational. The CS approach relies on the substitution of a component (obtained, e.g., by a spectral transformation of the data) of the multispectral (subsequently denoted as MS) image by the panchromatic (subsequently denoted as PAN) image. The CS class contains algorithms such as intensity-hue-saturation (IHS) [10, 11], principal component analysis (PCA) [12, 13]. The MRA approach is based on the injection of spatial details, which are obtained through a multiscale decomposition of the PAN image into the MS data. The spatial details can be extracted according to several modalities of MRA: decimated wavelet transform (DWT) [14], undecimated wavelet transform (UDWT) [15], Laplacian pyramid [16], nonseparable transforms, either based on wavelets (e.g., contourlets [17]) or not (e.g., curvelets [18]). The Bayesian approach relies on the use of posterior

distribution of the full resolution target image given the observed MS and PAN images. This posterior, which is the Bayesian inference engine, has two factors: (a) the likelihood function, which is the probability density of the observed MS and PAN images given the target image, and (b) the prior probability density of the target image, which promotes target images with desired properties, such as being spatially piecewise smooth. The selection of a suitable prior allows us to cope with the usual ill-contourlets of the pansharpening inverse problems. The variational class is interpretable as a particular case of the Bayesian one, where the target image is estimated by maximizing the posterior probability density of the full resolution image. The works [19–21] are representative of the Bayesian and variational classes.

With the increasing availability of HS systems, the pansharpening methods are now extending to the fusion of HS and panchromatic images [22, 23]. Pansharpening of HS images is still an open issue, and very few methods are presented in the literature to address it. Several of the methods designed for HS pansharpening were originally designed for the fusion of MS and HS data [24–30], the MS data constituting the high spatial resolution image.

However, a universal pansharpening technique does not yet exist, leaving end-users of the technology with an increasingly difficult task selecting a suitable approach. Challenges remain in: (1) co-registration of multisource data; (2) balance between spectral and spatial preservations; (3) high computational cost.

This section will introduce our recent hybrid method for hyperspectral image pansharpening, which is superior to the above challenges, with a specific application to fuse thermal hyperspectral and visible color images. Hybrid methods have been also proposed [28, 29] which use both component substitution and multiscale decomposition, such as guided filter and PCA.

6.2.1 Hybrid Method to Fuse Thermal Hyperspectral and Visible Color Images

Multisource data used in the experiments includes a thermal infrared (TI) hyperspectral data and a visual RGB image,[1] which were acquired by Telops Inc. on May 2013 over an urban area near Thetford Mines in Québec, Canada. The TI HS image has 84 spectral bands that cover the wavelengths between 7.8 to 11.5 μm with approximately 1-m spatial resolution. The visible RGB image is a series of color images acquired during separate flight-lines with approximately 20-cm spatial resolution. The whole scene of both data contains 7 classes, but with different spatial size of which the TI HS consists of 874 × 751 pixels while RGB of 4386 × 3769. Figure 6.2 shows an RGB composition with the labeled classes highlighted, for details, see [28].

Pansharpening thermal infrared HS data is much more complex than pansharpening general HS data. The spectral range of RGB/Pan and general HS images is typically within 0.4–2.5 μm, while thermal infrared HS image in the range 3.5–20 μm

[1] http://www.grss-ieee.org/community/technical-committees/data-fusion/2014-ieee-grss-data-fusion-contest/.

(e.g., Telops Inc. acquires thermal infrared HS data in the range 7.8–11.5 μm, which is out of the visible spectral range). Challenges remain for conventional pansharpening methods to yield good performances, where the original HS data and RGB/Pan image have no overlap spectral ranges.

6.2.1.1 Hybrid Fusion Method by Using Guided Filter in PCA Domain

One of the main challenges of the fusion of a low spatial resolution HS and a high-resolution RGB to get a high spatial resolution HS is to make a balance on spectral and spatial preservations. Recently, the guided filter [31] has been widely used in many applications (e.g., edge-aware smoothing, detail enhancement, etc.), as its efficient and strong abilities to transfer the structures of the guidance image to the filtering output.

The presented hybrid method [28, 29] enhances the spatial resolution of HS image by using guided filter [31] in PCA (principal component analysis) domain. We will introduce a two-stage hybrid pansharpening method [29] for HS image, Fig. 6.1 shows the corresponding flowchart (same in caption of Fig. 6.1). To fully exploit the spatial structures of HR RGB image and better co-register multisource data, in the first stage, we first downsample the HR RGB image to the same spatial resolution of LR HS image and use PCA to decorrelate the original LR HS images and separate the information content from the noise. The first k PCA channels contain most of the total energy of an HS image (i.e., most information of the HS image), and the remaining $B - k$ PCA channels (where B is the number of spectral bands of HSI and $B \gg k$) mainly contain noise. We use guided filter [31] to transfer the image details (e.g., edge) of the downsampled RGB image to the first k PCA channels only. We remove the noise (without guided filtering) in the remaining PCA channels using a soft-thresholding scheme. The improved LR HS image can be obtained by inverse PCA. In the second stage, we use the guided filter [31] in PCA domain to transfer the image details from the original HR RGB to the first few PCs of the improved LR HS image (we get from the first stage), same as we did in [28].

Let PC_i denote the i-th ($i \leq k$) PC of the original LR HS image, I'_{RGB} the RGB image downsampled by cubic interpolation to the same spatial resolution of the original LR HS image. The filtering output PC'_i can be represented as a linear transform of guided image I'_{RGB} in a local window ω of size $(2r + 1) \times (2r + 1)$ as follows:

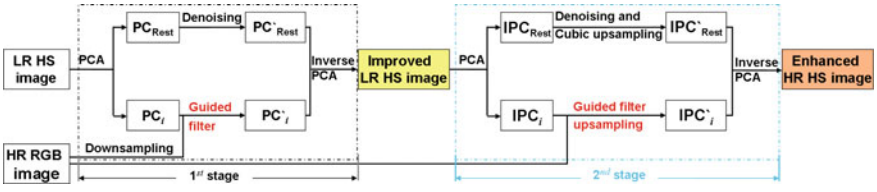


Fig. 6.1 Framework for hyperspectral image pansharpening using two-stage guided filter in PCA domain

$$PC'_i = a_j I'_{RGB} + b_j, \forall i \in \omega_j \tag{6.1}$$

The above model ensures that the output PC'_i has an edge only if the guided image I'_{RGB} has an edge, as $\nabla PC' = a\nabla I'_{RGB}$. The following cost function was used to find the coefficients a_j and b_j:

$$E(a_j, b_j) = \Sigma_{i\in\omega_j}((a_j I'_{RGB} + b_j - PC_i)^2 + \varepsilon a_j^2) \tag{6.2}$$

where ε is a regularization parameter determining the degree of the blurring for the guided filter. For more details about guided filter, we refer the readers to [31]. In the cost function, the $PC'_i = a_j I'_{RGB} + b_j$ should be as close as possible to the PC_i, which enforces the preservation of the original spectral information. The remaining $B - k$ PCA channels (where B is the number of spectral bands of HSI and $B \gg k$) mainly contain noise. If guided filtering is performed on these noisy and high-dimensional $B - k$ PCs, then it will amplify the noise of the data cube and cause high computational cost in processing the data, which is undesirable. Therefore, we remove the noise (and without guided filter) in the remaining PCA channels using a soft-thresholding scheme. After the first stage, we can see from Fig. 6.2 that the improved LR HS image becomes sharper and contains less noise than the original

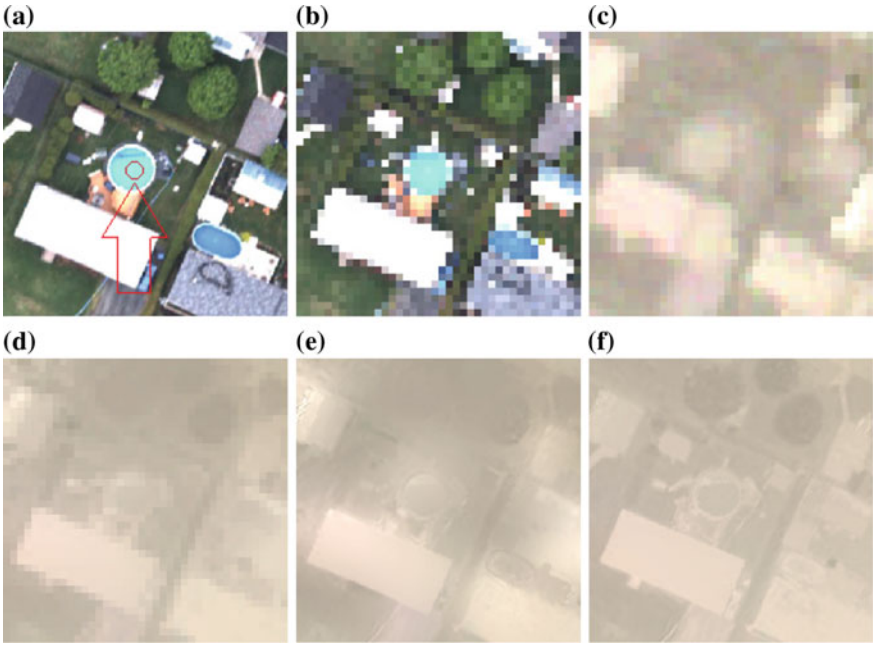


Fig. 6.2 Part of enhanced TI HS image. **a** HR visible RGB image; **b** downsampled RGB image to the same spatial resolution of TI HS image; **c** three bands composition by original LR HS image; **d** improved LR HS image (in the first stage); **e** enhanced HS image by GFPCA [28]; and **f** enhanced HS image by the two-stage hybrid fusion method

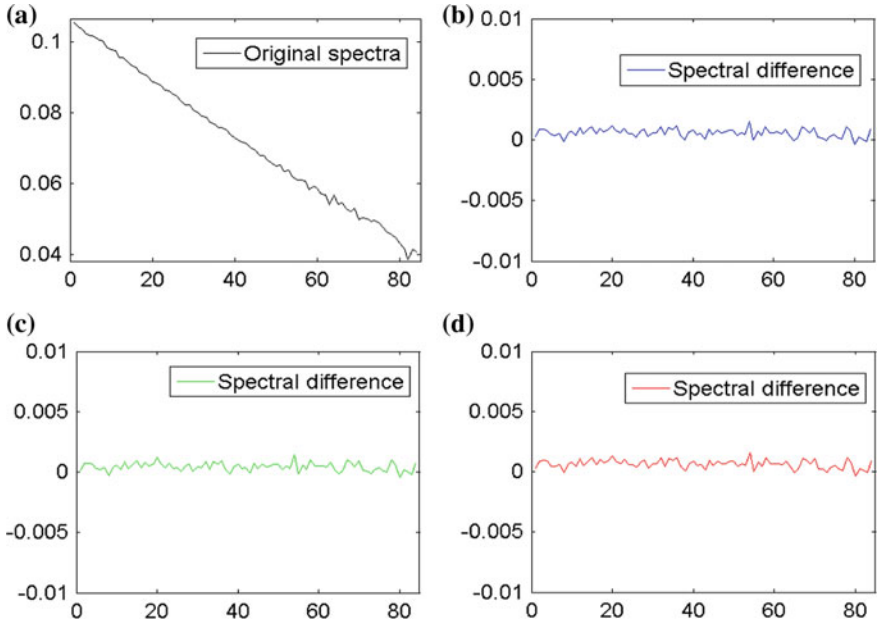


Fig. 6.3 Preservation of the spectral information. **a** The original spectra; and its difference with **b** improved LR HS image; **c** GFPCA [28]; and **d** the two-stage hybrid fusion method

LR HS image. Moreover, the first stage ensures better co-registration of multisource data for the following spatial enhancement.

In the second stage, we perform upsampling on the improved LR HS image from the first stage by combining the original HR RGB image and guided filter [31] in PCA domain, same as we did in [28]. For more details about our previous hyperspectral pansharpening method by guided filter in PCA domain, the readers can find relevant information in [9, 28]. Figures 6.2 and 6.3 show the effectiveness of the two-stage hybrid hyperspectral image pansharpening method in spectral and spatial preservations. From a visual analysis, the fused image produced by the two-stage hybrid fusion method appears to be sharper than our previous method [28], where only one stage of guided filtering was performed in PCA domain. Moreover, the spectral preservations of the two-stage hybrid fusion method and [28] are similar, and spectrally consistent with respect to the original HS image. With only one-stage fusion, GFPCA [28] mistook some big bright objects in the original LR HS as an object in the enhanced HS, leading to poor preservation of spatial information.

6.2.1.2 Experimental Results and Analysis

To evaluate the quality of the enhanced image products, here we use them for a specific application (e.g., classification). The proposed method is applicable to general HS

pansharpening tasks with no restrictions related to specific classification problems. However, typical image quality assessments require a reference image, which is very difficult to obtain in real applications. We used a support vector machines (SVM) [32] classifier, as it performs well even with a limited number of training samples, limiting the Hughes phenomenon. The SVM classifier with radial basis function (RBF) kernels in MATLAB SVM Toolbox, LIBSVM [33], is applied in our experiments. SVM with RBF kernels has two parameters: the penalty factor C and the RBF kernel widths γ. We apply a grid-search on C and γ using fivefold cross-validation to find the best C within the given set $\{10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}\}$ and the best γ within the given set $\{10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}\}$. We compare the two-stage hybrid (Hybrid2) hyperspectral pansharpening method with the schemes of (1) Simply enlarging the original HS image by cubic interpolation (Cub); (2) Only using the visible RGB image (RGB); (3) PCA component substitution method (PCA), similar as [13]; (4) Our previous method using guided filter in PCA domain (GFPCA) [28]. For quantitative comparisons, we randomly select 1000 samples per class from training set for training, the results are averaged over five runs. The classification results are quantitatively evaluated by measuring the Overall Accuracy (OA), the Average Accuracy (AA), and the Kappa coefficient (κ) on the test samples. The experiments were carried out on 64-b, 3.40 GHz Intel i7-4930K (1 core) CPU computer with 64 GB memory, the consumed time includes image fusion, feature fusion, and classification. Table 6.1 shows the accuracies and consumed time (hours) obtained from the experiments, Fig. 6.4 shows the best result of each method.

It is obvious that using single data source is not enough for reliable classification. By using only the spatial information from HR RGB image, we produce better results than simply upsampling the original HS image by cubic interpolation. However, the remote sensing data from urban area was a mix between man-made structures and natural materials, different objects may share similar spatial information. For example, the spatial information of 'Red roof' and 'Grey roof' or 'Bare soil' is similar, objects from 'Red roof' are misclassified as soil by only using RGB image. Image fusion by Cub cannot preserve the spatial information, leading to spatial distortions in the final classification map; whereas the PCA component substitution suffers from spectral distortions. By using the guided filter, the GFPCA performs better on both spectral and spatial preservations, and this is reflected in classification accuracy 20% higher than Cub and PCA, respectively.

Table 6.1 Average classification accuracies and consumed time (hour) obtained according to the described scheme

	Cub	RGB	PCA	GFPCA	Hybrid2
OA (%)	52.6	77.2	59.9	79.0	**84.3**
AA (%)	35.1	77.9	54.6	66.8	**78.5**
κ (%)	34.7	67.5	46.3	68.7	**76.7**
Time (hours)	3.81	**0.28**	2.82	1.17	1.23

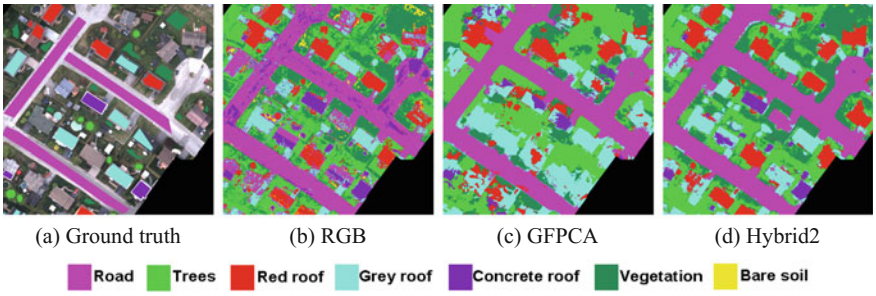


Fig. 6.4 Part of classification maps obtained by each method

By using the two-stage hybrid pansharpening method, we have more than 5% improvements in accuracies over the others fusion schemes. Although consuming a little bit more time than one-stage pansharpening method [28],the two-stage hybrid hyperspectral image pansharpening method benefits better preservations of image edges in its classification map.

6.3 Graph-Based Feature Fusion Model for Multisource Data Classification

Diverse sensor technologies and image processing algorithms allow us to measure different aspects of objects on the Earth: spectral characteristics in hyperspectral (HS) images, height in Light Detection And Ranging (LiDAR) data, geometry in image processing technologies like Morphological Profiles (MPs) (see also Chap. 7). It is obvious that no single technology can always suffice for reliable image interpretation. A key potential of multiple imaging sources is in their complementarity: each source can bring to the light some information that cannot be deduced from the other data sources. Hyperspectral (HS) images provide a detailed description of the spectral signatures of ground covers, whereas LiDAR data gives detailed information about the height of the same surveyed area. Multisource data, once combined, can provide a more comprehensive interpretation of objects on the ground. For example, HS spectral signatures could not be used to differentiate objects made with same material (e.g., roofs and roads made with the same asphalt), while LiDAR data can. On the other hand, LiDAR data alone could not be used to differentiate objects with same elevation (e.g., grassy areas and roads on the same flat surface), while HS spectral signatures can. Neither HS nor LiDAR data could separate man-made objects made with same material and with the same elevation (e.g., roads and parking lots), while they can often be easily distinguished by their geometry. Due to the increased availability of multisource data, the fusion of these remote sensing data has been of great interest for many practical applications [34–39, 39–48].

In the literature, feature- and decision-level fusion are two of the most popular methods that can be used to increase the capabilities of pattern recognition. Gunatilaka and Baertlein [49] compared the performances of feature-level and decision-level fusion algorithms for multisource data acquired by a metal detector, a ground-penetrating radar, and an infrared camera. They found that decision-level fusion does not always perform significantly better than the best available sensor. The performance of feature-level fusion is significantly better than the individual sensors. Feature-level fusion was reported to utilize the complementary information (from functional magnetic resonance imaging (fMRI), structural MRI (sMRI), and electroencephalography (EEG)) to provide additional insight into connectivity across brain networks and changes due to disease. For example, joint independent component analysis [50] was proposed to explore associations across multiple feature sources (generated from fMRI, sMRI, and EEG) through variations across individuals [51], providing new information about the brain. Canonical correlation analysis was exploited for feature fusion of fMRI and sMRI dataset (as well as fMRI and EEG data) [52], showing an interesting joint relationship between fMRI and gray matter. In the field of biometric which concerns about security, privacy, and forensics, the information integration at the feature-level reported more reliable recognition performances than other levels of fusion [53, 54]. For example, Nagar et al. [55] proposed a feature-level fusion framework based on fuzzy vault and fuzzy commitment to integrate multiple biometric data of fingerprint, iris, and face, reporting higher security and matching performance than their unibiometric counterparts to the application of cryptosystems.

For the tasks of remote sensing scene classification, Koetz et al. classified fuel composition from fused LiDAR and HS bands using Support Vector Machines (SVM) [36], showing the classification accuracies from fusion were higher than from either sensor alone. The joint use of HS and LiDAR remote sensing data for the classification of complex forest areas was investigated in [37]. They proposed a novel classification system, based on different possible classifiers that were able to properly integrate multisource information. In [35], Swatantrana et al. explored fusion of structural metrics from the LiDAR data and spectral characteristics from HS data for biomass estimation in the Sierra Nevada. Naidooa et al. [38] classified eight common savanna tree species in the Greater Kruger National Park region, South Africa, by fusing HS and LiDAR data in an automated Random Forest modeling approach. For multiple feature fusion, some of these approaches employ the so-called composite kernel methods [43, 44, 56–60] or their generalization [45]. Others define spatial information through morphological profiles and concatenate spectral and spatial features in a stacked architecture for classification [34, 39]. For example, Camps–Valls et al. [56] applied kernel methods to fuse different sources of information in a very natural way for a multitemporal image classification task. Multiple kernel learning was used in multimodal studies for combining spectral and spatial information [57], optical and radar data [43, 58], data form the same satellite but completely different locations [59], or optical data from different satellites [60]. These kernel methods typically generate a kernel matrix from each data source, and then fuse all the source-specific kernel matrices by linear combination into a multimodal similarity matrix.

Fauvel et al. in [34] concentrated multiple feature sources extracted from HS data for a classification task. The approach of [39] applied morphological attribute profiles (EAPs) [61] to both HS and LiDAR data for a classification task. Their method jointly considered the features extracted by EAPs computed on both HS and LiDAR data, and fused spectral, spatial, and elevation information in a stacked architecture.

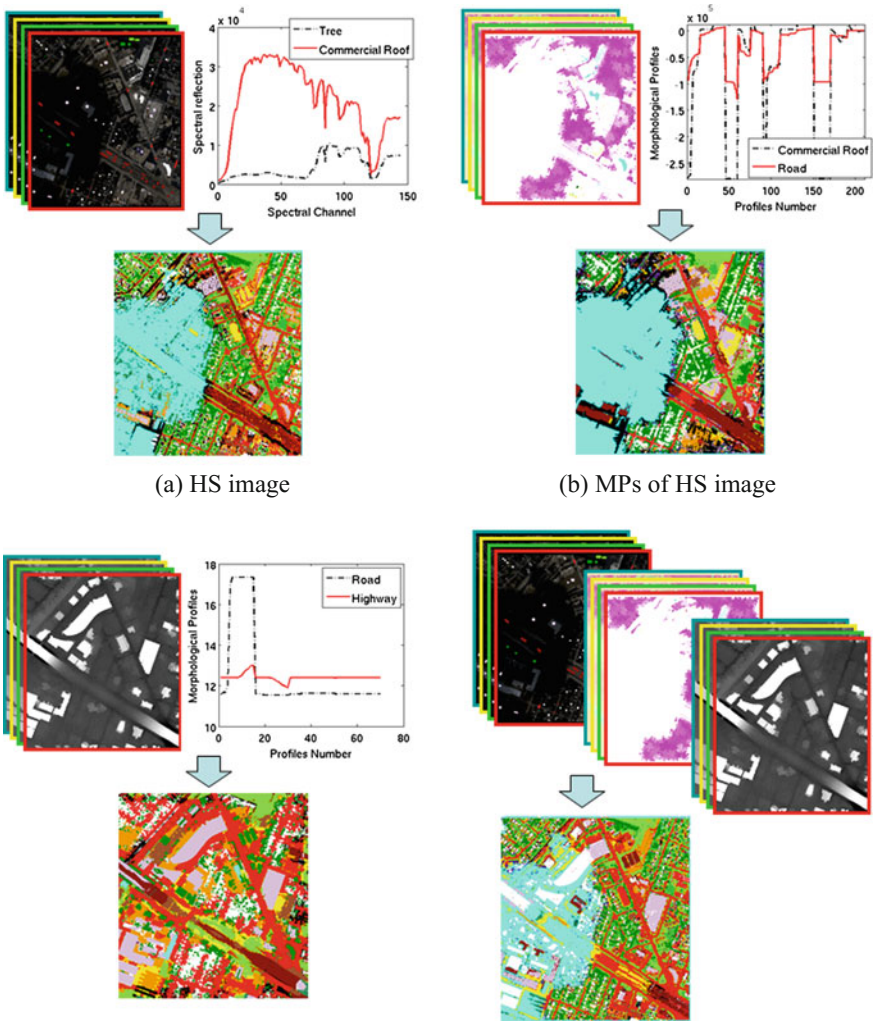
While such methods (that simply concatenate several kinds of features together) are appealing due to their simplicity, they may not perform better (or may even perform worse) than using a single feature, see Fig. 6.5. Dalla Mura et al. [62] showed examples where the classification accuracies after stacking different morphological attributes even dropped compared to the case of considering a single one. This is because the information contained by different features is not equally represented or measured. The value of different components in the feature vector can be significantly unbalanced. Furthermore, stacking several kinds of feature sources may yield redundant information. In addition, the increase in the dimensionality of the stacked features, as well as the limited number of labeled samples may in practice pose the problem of the "curse of dimensionality," consequently increasing the risk to overfit the training data.

Recently, the graph-based fusion method of [42, 63] fuses multisource data using a combination of morphological and spectral features. Markov modeling formalizes spatial and multimodal fusion through global minimum energy concepts [46]. Domain adaptation and manifold alignment [47] combine multisource data at a geometrical level in a latent space, regardless of the different nature and dimensionality of the sources. Sparse dictionary learning conducts fusion at the signal or information-theoretic level [48]. Multiple feature learning was applied to integrate multiple morphological features generated from both HS and LiDAR data [40]. An adaptive joint sparse representation classification model was presented for fusion of multisource remote sensing data in [41].

In this section, we will present graph-based feature-level fusion model to couple dimensionality reduction and data fusion of multisource data for a classification task, where the advantages of each data source are considered.

6.3.1 Graph Fusion of Hyperspectral and LiDAR Data

In this subsection, a graph-based fusion method is presented to couple dimensionality reduction and data fusion of multiple feature sources (generated from both HS and LiDAR data) together for a classification task. First, morphological operations are used to model spatial and elevation information from HS and LiDAR data, respectively. Then, a fusion graph is built where only the dimensional normalized feature points with similar spectral, spatial, and elevation characteristics are connected. Finally, the problem of multisource data fusion is solved by projecting all the features into a low-dimensional subspace, on which neighborhood relationships among data points (i.e., with similar spectral, spatial, and elevation characteristics) in the original space are maintained.



(a) HS image (b) MPs of HS image

(c) MPs of LiDAR data (d) Fusion by simply stacking all feature sources

Fig. 6.5 Classification by using multisource data. Each data source has its advantages, fusion of multisource data by simply stacking them together does not always perform better than using single source

6.3.1.1 Morphological Features

Feature extraction is a critical step for multisource data fusion. With advanced modern imagers (e.g., hyperspectral), the resulting spatial, spectral, and temporal resolutions are intractably high. Feature extraction aims at reducing the information redundancy while preserving the most important information of the original data. Auto-

mated information extraction employs either object-based or pixel-based approaches. Object-based methods first group the image pixels in a meaningful way via image segmentation [64]. This approach provides a natural way to incorporate geometrical information by calculating different shape characteristics of the segmented objects. However, the segmentation process typically relies on parameters that are highly dependent on the image data at hand and on the specific tasks [64]. Pixel-based approaches often employ mathematical morphology [65] ranging from, low-level feature extraction (size and shape features) using morphological profiles [66], over middle-level attribute profiles [67] to high-level feature extraction with semantic information indexes [68]. Recent works demonstrate benefits of using mathematical morphology in modeling and extracting geometrical information from remote sensing images for classification, change detection, urban planning, and risk assessments [67–70].

Morphological features are generated by either applying morphological openings or closings by reconstruction on the image, using a structural element (SE) of predefined size and shape. For example, the morphological profile (MP) with disk SE carries information about the minimum size of objects, whereas directional MP indicates the maximum size of objects [71, 72]. An opening acts on bright objects (for LiDAR data, the bright regions are actually areas with the high elevation, such as the top of the roof) compared with their surrounding, while closings act on dark (low height in the LiDAR data) objects. For example, an opening deletes bright objects that are smaller than the SE. By increasing the size of the SE and repeating the previous operation, a complete morphological profile (MP) is built, carrying information about the size and the shape of objects in the image. More details on morphological operators will be presented in Chap. 7.

In our experiments, morphological features are generated by applying morphological openings and closings with partial reconstruction [71, 72] on both LiDAR data and the first 2 principal components (PCs) (representing more than 99% of the cumulative variance) of original HS image. The effect of using morphological features with partial reconstruction for classification of remote sensing data from urban areas has been discussed in our previous work [71, 72]. For disk-shaped SE, MPs with 15 openings and closings (ranging from 1 to 15 with step size increment of 1) are computed for both LiDAR data and the first 2 PCs of HS image. For linear structuring elements, MPs with 20 openings and closings (ranging from 5 to 100 with step size increment of 5) are constructed for both LiDAR data and the first 2 PCs of HS image. Figures 6.6 and 6.7 show the results of MP with partial reconstruction for both LiDAR data and the first PC of HS image in different scales. As the size of the SE increases in openings, we can see that more and more bright objects (i.e., objects with high elevation) disappear in the dark background of LiDAR data. More and more dark objects disappear in the closings of the first PC of HS image.

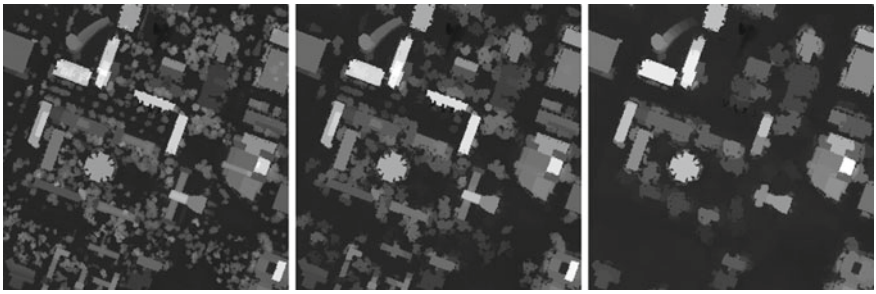


Fig. 6.6 Openings on a part of LiDAR data with disk-shaped SEs of increasing size (1, 3, and 5)

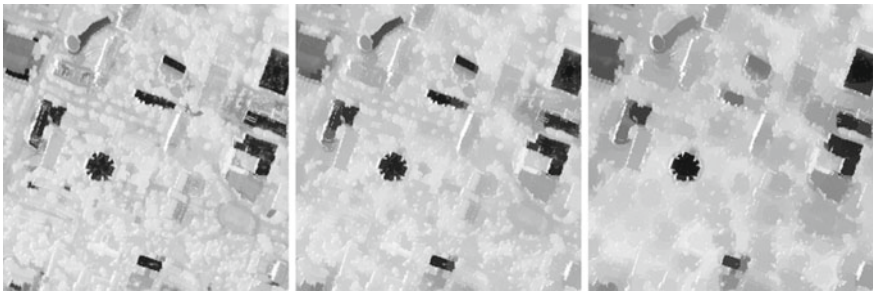


Fig. 6.7 Closings on a part of the 1st PC of the HS image with disk-shaped SEs of increasing size (1, 3, and 5)

6.3.1.2 Feature Dimension Normalization

Different features may have different dimensionalities and characteristics. The differences in feature space and scaling make it difficult to homogenize features from multiple sources. For example, in our experiments, the original hyperspectral image with 144 bands contains the spectral information of the ground covers. The morphological features of LiDAR data with 70 bands (with 30 bands disk-based MP and 40 bands directional MP) carry the elevation information of the same surveyed area, and the morphological features of HS image with 140 bands contain the spatial information. Before fusing the features, we first need to normalize the feature dimension and reduce the computational cost and the noise throughout the given feature space. An effective way is to use Kernel Principal Component Analysis [73] for dimensionality reduction on each feature separately. The normalized dimension of each feature space can be chosen as the smallest dimension of all these features. Without losing generality, in this paper, we assume the dimension of each feature is already normalized to $D = 70$.

6.3.1.3 Graph-Based Feature Fusion Method

Let $\mathbf{X}^{Spe} = \{\mathbf{x}_i^{Spe}\}_{i=1}^n, \mathbf{X}^{Spa} = \{\mathbf{x}_i^{Spa}\}_{i=1}^n$ and $\mathbf{X}^{Ele} = \{\mathbf{x}_i^{Ele}\}_{i=1}^n$, where $\mathbf{x}_i^{Spe} \in \mathbb{R}^D, \mathbf{x}_i^{Spa} \in \mathbb{R}^D$ and $\mathbf{x}_i^{Ele} \in \mathbb{R}^D$ denote the spectral, spatial, and elevation features, respectively,

after normalization to the same dimension. $\mathbf{X}^{Sta} = \{\mathbf{x}_i^{Sta}\}_{i=1}^n = [\mathbf{X}^{Spe}; \mathbf{X}^{Spa}; \mathbf{X}^{Ele}]$ and $\mathbf{x}_i^{Sta} = [\mathbf{x}_i^{Spe}; \mathbf{x}_i^{Spa}; \mathbf{x}_i^{Ele}] \in \mathbb{R}^{3D}$ denote the vector stacked by the spectral, spatial, and altitude features. $\{\mathbf{z}_i\}_{i=1}^n$, and $\mathbf{z}_i \in \mathbb{R}^d$ denote the fusion features in a lower dimensional feature space with $d \leq 3D$. The goal of this paper is to find a transformation matrix $\mathbf{W} \in \mathbb{R}^{3D \times d}$, which can couple dimensionality reduction and feature fusion in a way of $\mathbf{z}_i = \mathbf{W}^T \mathbf{x}_i$ ($\mathbf{x}_i$ is a variable, which can set to be $\mathbf{x}_i^{Sta}$, $\mathbf{x}_i^{Spe}$, etc.). The transformation matrix $\mathbf{W}$ should not only fuse different features in a lower dimensional feature space, but also preserve local neighborhood information and detect the manifold embedded in the high-dimensional feature space. A reasonable way [74] to find the transformation matrix $\mathbf{W}$ can be defined as follows:

$$\arg \min_{\mathbf{W} \in \mathbb{R}^{3D \times d}} \left(\sum_{i,j=1}^{n} ||\mathbf{W}^T \mathbf{x}_i - \mathbf{W}^T \mathbf{x}_j||^2 A_{ij} \right) \tag{6.3}$$

where the matrix $\mathbf{A}$ encodes the edges of a graph $\mathbf{G}$ with nodes $\mathbf{X}$. We assume that the edge (between data point $\mathbf{x}_i$ and $\mathbf{x}_j$) $A_{ij} \in \{0, 1\}$; $A_{ij} = 1$ if $\mathbf{x}_i$ and $\mathbf{x}_j$ are "close" and $A_{ij} = 0$ if $\mathbf{x}_i$ and $\mathbf{x}_j$ are "far apart". The "close" here is defined by finding the k nearest neighbors (kNN) of the data point $\mathbf{x}_i$. The kNN is determined first by calculating the distance (we use Euclidean distance here) between data point $\mathbf{x}_i$ and all the data points, then sorting the distance and determining nearest neighbors based on the k-th minimum distance. Minimizing the objective function of Eq. (6.3) ensures that if $\mathbf{x}_i$ and $\mathbf{x}_j$ are "close" then $\mathbf{z}_i$ and $\mathbf{z}_j$ are close as well.

When the graph is constructed by spectral features (i.e., $\mathbf{G} = \mathbf{G}^{Spe} = (\mathbf{X}^{Spe}, \mathbf{A}^{Spe})$), the k nearest neighbors (i.e., $A_{ij}^{Sta} = 1$, $j \in \{1, 2, \cdots, k\}$) of the data point $\mathbf{x}_i^{Spe}$ indicate the spectral signatures of these kNN data points $\mathbf{x}_j^{Spe}$, which are the most similar in terms of Euclidean distance. We propose a fusion graph which we define $\mathbf{G}^{Fus} = (\mathbf{X}^{Sta}, \mathbf{A}^{Fus})$ as follows:

$$\mathbf{A}^{Fus} = \mathbf{A}^{Spe} \odot \mathbf{A}^{Spa} \odot \mathbf{A}^{Ele} \tag{6.4}$$

where the operator '$\odot$' denotes element-wise multiplication, i.e., $A_{ij}^{Fus} = A_{ij}^{Spe} A_{ij}^{Spa} A_{ij}^{Ele}$. Note that $A_{ij}^{Fus} = 1$ only if $A_{ij}^{Spe} = 1$, $A_{ij}^{Spa} = 1$ and $A_{ij}^{Ele} = 1$. This means that the stacked data point $\mathbf{x}_i^{Sta}$ is "close" to $\mathbf{x}_j^{Sta}$ only if all individual feature points $\mathbf{x}_i^{Ind}$ ($Ind \in \{Spe, Spa, Ele\}$) is "close" to $\mathbf{x}_j^{Ind}$. The connected data points $\mathbf{x}_i^{Sta}$ and $\mathbf{x}_j^{Sta}$ have similar spectral, spatial, and altitude characteristics. If any individual feature point $\mathbf{x}_i^{Ind}$ is "far apart" from $\mathbf{x}_j^{Ind}$, then $A_{ij}^{Fus} = 0$. In real data, the data points from the football fields made by real grass ($\mathbf{x}_i^{Sta}$) and by synthetic grass ($\mathbf{x}_j^{Sta}$) have similar spatial and altitude information ($\mathbf{A}_{ij}^{Spa} = 1$, $\mathbf{A}_{ij}^{Ele} = 1$), but different spectral characteristics ($\mathbf{A}_{ij}^{Spe} = 0$), so these two data points are not "close" (i.e., $\mathbf{A}_{ij}^{Fus} = 0$). By reformulating the objective function of Eq. (6.3), we obtain:

$$\begin{aligned}\sum_{i,j=1}^{n} ||\mathbf{W}^T\mathbf{x}_i^{Sta} - \mathbf{W}^T\mathbf{x}_j^{Sta}||^2 A_{ij} &= 2\left(\sum_{i=1}^{n}\mathbf{W}^T\mathbf{x}_i^{Sta}D_{ii}(\mathbf{x}_i^{Sta})^T\mathbf{W} - \sum_{i,j=1}^{n}\mathbf{W}^T\mathbf{x}_i^{Sta}A_{ij}(\mathbf{x}_j^{Sta})^T\mathbf{W}\right)\\ &= 2\mathbf{W}^T\mathbf{X}^{Sta}(\mathbf{D}-\mathbf{A}^{Fus})(\mathbf{X}^{Sta})^T\mathbf{W}\\ &= 2\mathbf{W}^T\mathbf{X}^{Sta}\mathbf{L}^{Fus}(\mathbf{X}^{Sta})^T\mathbf{W}\end{aligned}$$

where $\mathbf{D}$ is a diagonal matrix with $D_{ii} = \sum_{j=1}^{n} A_{ij}^{Fus}$, $\mathbf{L}^{Fus} = \mathbf{D} - \mathbf{A}^{Fus}$ is the fusion Laplacian matrix. Larger D_{ii} (corresponding to $\mathbf{z}_i$) indicates more important of $\mathbf{z}_i$. To avoid degeneracy, we impose a constraint, similar as [75]:

$$\mathbf{z}^T\mathbf{D}\mathbf{z} = 1 \Rightarrow \mathbf{W}^T(\mathbf{X}^{Sta})\mathbf{D}(\mathbf{X}^{Sta})^T\mathbf{W} = \mathbf{I}$$

Therefore, the objective function of Eq. (6.3) with the imposed constraint is:

$$\begin{aligned}\underset{\mathbf{W}}{\text{minimize}} \quad & \mathbf{W}^T\mathbf{X}^{Sta}\mathbf{L}^{Fus}(\mathbf{X}^{Sta})^T\mathbf{W}\\ \text{subject to} \quad & \mathbf{W}^T(\mathbf{X}^{Sta})\mathbf{D}(\mathbf{X}^{Sta})^T\mathbf{W} = \mathbf{I}\end{aligned} \tag{6.5}$$

Finally, we can obtain the transformation matrix $\mathbf{W} = (\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_r)$ which is made up by r eigenvectors associated with the least r eigenvalues $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_r$ of the following generalized eigenvalue problem:

$$(\mathbf{X}^{Sta})\mathbf{L}^{Fus}(\mathbf{X}^{Sta})^T\mathbf{w} = \lambda(\mathbf{X}^{Sta})\mathbf{D}(\mathbf{X}^{Sta})^T\mathbf{w}. \tag{6.6}$$

6.3.1.4 Experimental Results and Analysis

To assess the quality of the fusion products, we here use them for a specific classification task. Experiments are done on a hyperspectral image and a LiDAR data set which were acquired by the NSF-funded Center for Airborne Laser Mapping (NCALM) on June 2012 over the University of Houston campus and the neighboring urban area. The hyperspectral imagery has 144 spectral bands with wavelength range from 380 to 1050 nm. Both datasets have the same spatial resolution (2.5m). The whole scene of the data,[2] consisting of the full 349 × 1905 pixels, contains 15 classes. Available training and testing set are given in Table 6.2, and Fig. 6.8 shows false color image of HS data and LiDAR image. Note that the color in the cell of Table 6.2 denotes different classes in the classification maps in Fig. 6.8. For more information, please refer to [76].

We keep the same settings for SVM classifier as in Sect. 6.2.1.2, and compare graph-based data fusion method (GDF) with the schemes of (1) Using raw HS image (RAW_H); (2) Using the MPs computed on the first 2 PCs of original HSI (MPs_H); (3) Using the MPs computed on the LiDAR data (MPs_L); (4) Stacking morphological features computed from both LiDAR data and the first 2 PCs of original HS image

[2] http://www.grss-ieee.org/community/technical-committees/data-fusion/2013-ieee-grss-data-fusion-contest/.

Table 6.2 Training and test samples for multisource data on Houston University areas

Class Name	# Training set	# Test set	Class Name	# Training set	# Test set	Class Name	# Training set	# Test set
Grass Healthy	198	1053	Grass Stressed	190	1064	Grass Synthetis	192	505
Tree	188	1056	Soil	186	1056	Water	182	143
Residential	196	1072	Commercial	191	1053	Road	193	1059
Highway	191	1036	Railway	181	1054	Parking Lot 1	192	1041
Parking Lot 2	184	285	Tennis Court	181	247	Running Track	187	473

(MPs$_{H+L}$), similarly as [39]; (5) Stacking all dimensional normalized features, i.e., $\mathbf{X}^{Sta}$, we call it STA; (6) Stacking all the features extracted by PCA from each individual features which represents more than 99% of the cumulative variance (PCA); (7) Stacking all the features extracted by nonparametric weighted feature extraction (NWFE) [77] from each individual feature, as [34] fused the spectral and spatial information (NWFE); (8) Features fused by using the graph constructed by stacked features $\mathbf{X}^{Sta}$(i.e., locality preserving projection [75]) (LPP).

The classification results are quantitatively evaluated by measuring the Overall Accuracy (OA), the Average Accuracy (AA), and the Kappa coefficient (κ) on the test samples. Table 6.3 shows the accuracies obtained from the experiments. For visual comparison, we show the classification maps in Fig. 6.8.

Experimental results show that the objects in the cloud-covered regions are not well classified by only using HS data or the MPs of HS data, see the bottom of both the false color composite image of HS image and the classification maps in Fig. 6.8. Using the MPs of LiDAR performs better in these regions, see the accuracies of class 'Commercial' and 'Highway' in Table 6.3, most test samples of these two classes are in these cloud-covered regions. The elevation features alone are not enough to differentiate objects with same elevation (e.g., grassy areas and roads on the same flat surface), see the accuracies of class 'Grass Healthy,' 'Road,' and 'Running Track' in Table 6.3. By stacking morphological features computed both from LiDAR data and HS data, the overall accuracies are improved. The graph-based feature fusion method performed the best, with more than 8–20% improvements compared to the results of only using single features, and with 2–5% improvements with respect to the other fusion schemes. The schemes of (6) and (7) are similar to (5) in terms of a stacked architecture. The differences are that each individual feature is represented by different aspects, e.g., the features extracted by PCA represent most of the cumulative variance in the data, while the features extracted by NWFE respect the class discriminant. The cloud-covered regions in original HS image are not classified well by fusing features in a stacked architecture, the schemes of (5), (6), and (7) produced lower accuracies of 'Commercial' than only using the MPs of LiDAR, because the elevation information contained in the morphological features of LiDAR data is not well represented in such a way of data fusion. The spectral and the spatial information of the cloud-covered regions are not related to real ground cover. The LiDAR sensor can penetrate clouds and its morphological features contain the elevation information of the real ground cover in this cloud-covered region. When stacking all features together, the element values of different features can be significantly unbalanced, and

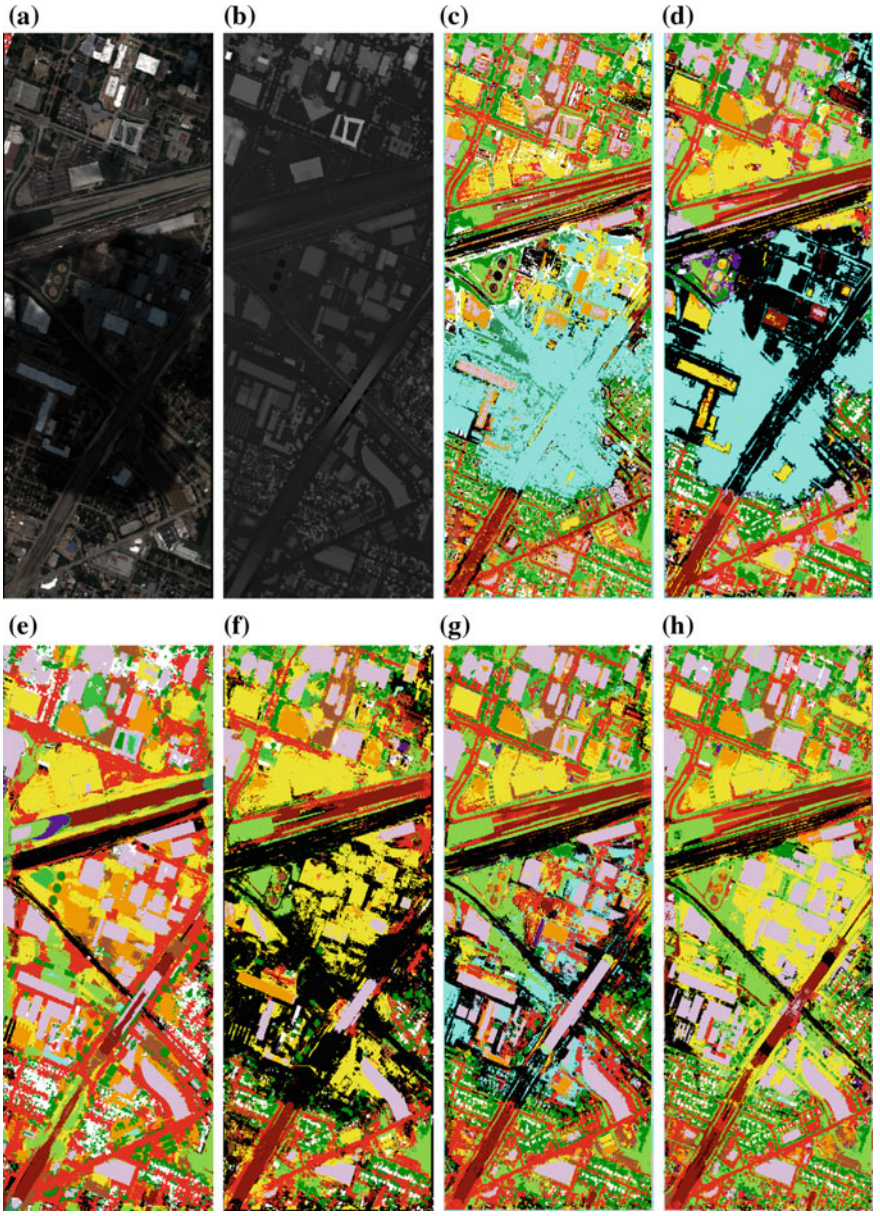


Fig. 6.8 Classification maps (part) produced by the described schemes. **a** False color image of HS data; **b** LiDAR data; and thematic maps using **c** raw HS data; **d** MPs of HS data; **e** MPs of LiDAR data; **f** the stacked features $\mathbf{X}^{Sta}$; **g** the features fused by LPP on $\mathbf{X}^{Sta}$; **h** graph-based data fusion method

Table 6.3 Classification accuracies for multisource data obtained by the described schemes on Houston University areas

No. of features	RAW$_H$	MPS$_H$	MPS$_L$	MPS$_{H+L}$	STA	PCA	NWFE	LPP	GDF
	144	140	70	210	210	35	42	26	26
OA (%)	80.72	82.43	69.39	86.39	87.49	85.28	87.96	87.81	**90.30**
AA (%)	83.40	84.99	68.42	88.48	88.94	87.29	88.76	88.88	**91.30**
κ (%)	79.23	81.02	66.79	85.31	86.42	84.02	86.92	86.80	**89.48**
Grass healthy	82.15	80.25	35.61	**82.43**	81.10	78.63	81.29	81.10	73.31
Grass stressed	81.58	80.64	67.11	82.61	84.87	81.77	83.27	82.80	**97.84**
Grass synthetis	99.80	**100.00**	79.60	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
Tree	92.80	84.09	72.92	91.10	95.45	93.75	89.49	97.73	**97.82**
Soil	97.92	**100.00**	83.52	99.91	99.91	99.91	99.81	98.77	99.24
Water	95.10	95.10	66.43	**100.00**	95.80	95.80	95.80	95.10	99.30
Residential	76.21	87.31	76.59	80.97	86.94	84.70	86.38	84.24	**88.15**
Commercial	54.51	45.58	91.45	63.06	59.54	66.95	76.07	79.20	**96.20**
Road	78.47	91.03	59.21	91.88	90.37	83.66	**93.58**	91.41	86.59
Highway	60.04	60.42	64.86	64.67	65.44	57.53	62.16	61.49	**76.83**
Railway	79.51	87.10	88.24	93.45	**99.24**	97.34	98.39	92.51	92.41
Parking lot 1	82.90	86.84	70.89	97.89	99.33	91.74	**99.90**	72.98	85.69
Parking lot 2	72.63	76.49	55.09	**79.30**	77.19	77.54	65.26	76	76.49
Tennis court	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
Running track	97.25	**100.00**	14.80	**100.00**	98.94	**100.00**	**100.00**	97.25	99.58

the information contained by different features is not equally represented. The same problems happen when using the stacked features to build a graph in LPP method. By considering the graph-based fusion method, better accuracies are obtained, and the cloud-covered regions of HS image are better classified.

6.3.2 *Local Graph Fusion Model for Fusion of Multisource Data*

The above graph-based data fusion method (GDF) [76] proved to overcome the conventional approach of stacking different feature sources together in terms of classification accuracy. The effectiveness of using such graph to fuse multiple feature sources for classification has been discussed in the very recent studies [42, 76]. However, the GDF can cause some problems on storage resources and computational load especially when using large training data sets. This is because finding k nearest neighbors to build a graph is very intensive in both computation and memory consumption. Random sampling was used to speed up the GDF in [76]. However, random sampling can lead to poor representation of the whole area if large areas are not sampled, which will lead to unstable performances. This is even worse if the

study area is very large and the number of samples fixed. Moreover, image degradation cannot be avoided during the hyperspectral data acquisition, which will lead to poor performances on finding k nearest neighbors for building a graph on the whole original data or randomly selected samples.

To overcome the above-mentioned limitations, this subsection will focus on local graph fusion of multisource data. Specifically, the local graph fusion model (LGF) [63] is used to couple dimensionality reduction and the fusion of multisource features (e.g., the original spectral feature and spatial features computed from the HS image). The main contributions of the local graph fusion model can be summarized as follows:

- First and foremost, the local graph fusion model builds the local fusion graph on the whole data by employing a sliding window. This way we introduce a different approach with regard to GDF [76], where the fusion graph was built globally on randomly selected samples. The local spatial neighborhood information is very important for remote sensing, especially for high-resolution remote sensing imagery. Specifically, many methods [78–81] demonstrated notable improvements on the performances of dimensionality reduction, classification, and segmentation, by exploiting the local spatial neighborhood information. In typical remote sensing scenes (especially for high-resolution remote sensing images), pixels in a small spatial neighborhood usually share similar properties (e.g., very similar spectral characteristics). If we build a fusion graph globally, pixels from different objects may become the nearest neighbors of each other, if they share similar spectral characteristics. For example, pixels belonging to a roof of a building may get connected in the graph to pixels of parking lots, because they have very similar spectral characteristics even though they might not be spatially adjacent. Within a small spatial window, LGF better employs the local spatial neighborhood information to represent objects in the feature space. This way, local graph fusion model enables better performances on classification, and better constraints in terms of local connectivity reduce a risk of erroneously selected nearest neighbors even when the spectral characteristics are affected by noise.
- In addition, the local fusion graph reduces computational complexity from $O(N^2)$, which holds for the global fusion graph on the whole data to only $O(NS^2)$, where N denotes the total number of spatial pixels, $S \ll N$ is the size of the sliding window.
- Last but not least, the local graph fusion model admits a fast implementation by just spatially downsampling the original data, while keeping the performances stable. As shown in the experiments, for the high-resolution remote sensing images, spatially downsampling will not affect much the main spatial structure of the objects (i.e., leading to similar classification performances obtained without subsampling), but can efficiently reduce the computational complexity by a factor equal to the square of the spatial downsampling ratio.

Suppose $\mathbf{X}^{Spe} = \{\mathbf{x}_i^{Spe}\}_{i=1}^N$ and $\mathbf{X}^{Spa} = \{\mathbf{x}_i^{Spa}\}_{i=1}^N$ denote the spectral and spatial features after normalization of their values to the same interval (e.g., [0,1]), where $\mathbf{x}_i^{Spe} \in \mathbb{R}^B$, with B the number of bands and $\mathbf{x}_i^{Spa} \in \mathbb{R}^D$ (with $D = p(2M + 1)$ being generated by an EMP built on p PCs and with M filters), and N is the total number

of spatial pixels in a HS image. Further on, we denote the stacked spectral and spatial features by $\mathbf{X}^{Sta} = \{\mathbf{x}_i^{Sta}\}_{i=1}^{N} = [\mathbf{X}^{Spe}; \mathbf{X}^{Spa}]$, where $\mathbf{x}_i^{Sta} = [\mathbf{x}_i^{Spe}; \mathbf{x}_i^{Spa}] \in \mathbb{R}^{B+D}$.

For calculating the pairwise distance matrix to find kNN for constructing the graph, the complexity on storing the data and computational time are of $O(N^2)$ and $O(BN^2)$.[3] In this case, even in conventional remote sensing images, the pairwise distance matrix will exceed the memory capacity of ordinary personal computer. For example, an image of $N = 512 \times 512$ pixels, the size of the distance matrix is $N \times N = (512 \times 512) \times (512 \times 512)$ elements. Therefore, in GDF [76], a small number of samples (e.g., $n = 5000$) was selected from the whole original data to build the global graph. However, random sampling may not always well represent the full data, especially for the data with large study area, which will lead to unstable performances of the global fusion graph. Moreover, image degradation cannot be avoided during the data acquisition, which will lead to poor performances on finding kNN globally from the randomly selected samples to build global fusion graph.

The local graph-based fusion model (LGF) probes an image with a $S \times S$ sliding window, calculates the kNN of the current pixel considering the neighboring samples included by the window, and builds the fusion graph within this sliding window. Figure 6.9 illustrates an example considering a 7×7 sliding window centered at one pixel $\mathbf{x}_i^{Spe}$. This way we reduce the computational complexity of calculating pairwise distance matrix to $O(BNS^2)$ $(S \ll N)$, as well as a significant reduction in memory use.

Here, we leverage the fact that pixels within a spatial neighborhood are likely to share similar properties. This assumption is particularly valid when dealing with images of very high spatial resolutions. If we consider the spectral features (original HS image), we define the "spectral neighbors" within a spatial neighborhood of a pixel with spectrum $\mathbf{x}_i^{Spec}$ as those k pixels whose values are closest to it in terms of spectral signatures (i.e., nearest neighbors). Let $\mathcal{N}_k^{Spec}(i)$ denote the set of values of the k nearest neighbors of $\mathbf{x}_i^{Spec}$ within the neighborhood, $\mathcal{N}_k^{Spec}(i) = \{\mathbf{x}_{im}^{Spec}\}_{m=1}^{k}$

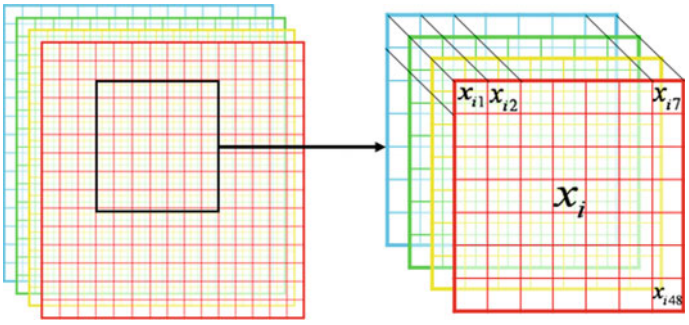


Fig. 6.9 An illustration of the 7×7 sliding window centered at pixel $\mathbf{x}_i$

[3] With faster algorithms (e.g., K-D trees) than direct nearest neighbours searching, the complexity can be reduced. More details on efficient representation and search techniques for large data sets can be found in Chap. 2.

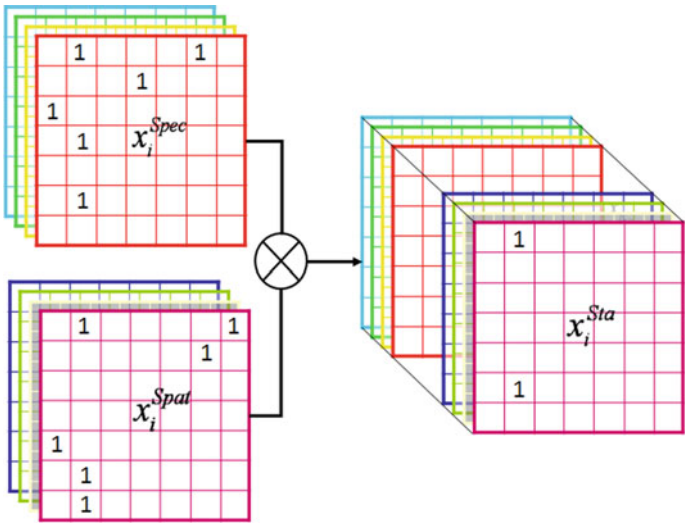


Fig. 6.10 An illustration of local fusion graph building within a 7×7 sliding window centered at pixel. The value '1' means connection (i.e., $E_{im} = 1$), while the blank grid means the data point $\mathbf{x}_{im}$ is not in the kNN of the current pixel $\mathbf{x}_i$ (i.e., $E_{im} = 0$)

and $m \neq i$, $|\mathcal{N}_k^{Spec}(i)| = k$ ($|\cdot|$ being the cardinality of the set). Then the edges in the graph $E_{im}^{Spec} = 1$ for $m \in \mathcal{N}_k^{Spec}(i)$ and $E_{im}^{Spec} = 0$ otherwise, $m \in \{1, \cdots, S^2\}$. Similarly, for the spatial features (i.e., EMP built on the HS image), the k nearest neighbors of the data point with values $\mathbf{x}_i^{Spat}$ are those of the neighborhood that are most similar to it in terms of the spatial characteristics. Analogously, $\mathcal{N}_k^{Spat}(i)$ denotes the k nearest spatial neighbors of $\mathbf{x}_i^{Spat}$, $|\mathcal{N}_k^{Spat}(i)| = k$, and thus $E_{im}^{Spat} = 1$ if $m \in \mathcal{N}_k^{Spat}(i)$ and $E_{im}^{Spat} = 0$ otherwise. LGF [63] constructs the fused kNN for the stacked features $\mathbf{X}^{Sta}$ within a spatial window as follows:

$$\mathcal{N}_k^{Fus}(i) = \mathcal{N}_k^{Spec}(i) \cap \mathcal{N}_k^{Spat}(i) \tag{6.7}$$

where the operator '$\cap$' denotes the intersection, i.e., the kNN of the stacked vector $\mathbf{x}_i^{Sta}$: $\mathcal{N}^{Fus}(i) = \{\mathbf{x}_{im}^{Sta}, m \in \mathcal{N}_k^{Spec}(i) \wedge m \in \mathcal{N}_k^{Spat}(i)\}$. The fused edge $\mathbf{E}_i^{Fus}$ for the stacked data point $\mathbf{x}_i^{Sta}$ must satisfy:

$$E_{i,m}^{Fus} = 1, \quad \text{iff} \quad m \in \mathcal{N}_k^{Spec}(i) \wedge m \in \mathcal{N}_k^{Spat}(i) \tag{6.8}$$

For instance, within the 7×7 sliding window centered at pixel $\mathbf{x}_i$ (when the sliding window is close to the image boundary, the symmetric padding is utilized to deal with the margin effect [82]), suppose the 6 nearest neighbors of spectral feature point $\mathbf{x}_i^{Spec}$ is $\mathcal{N}_6^{Spec}(i) = \{\mathbf{x}_{im}^{Spec} : m \in [2, 6, 11, 15, 23, 36]\}$, see Fig. 6.10. While the 6 nearest neighbors of spatial feature point $\mathbf{x}_i^{Spat}$ is $\mathcal{N}_6^{Spat}(i) = \{\mathbf{x}_{im}^{Spat}$:

$m \in [2, 7, 13, 28, 36]\}$. Therefore, we can get the kNN of fusion graph $\mathcal{N}_6^{Fus}(i) = \{\mathbf{x}_{im}^{Fus} : m \in [2, 36]\}$ according to Eq. (6.7). Then, we set their corresponding edges $E_{im}^{Fus} = 1$, for $m \in \mathcal{N}_6^{Fus}(i)$; $E_{im}^{Fus} = 0$ if $m \notin \mathcal{N}_6^{Fus}(i)$, $1 \leqslant m \leqslant 49$.

This means that the stacked data point $\mathbf{x}_i^{Sta}$ is "close" to $\mathbf{x}_{im}^{Sta}$ only if they have similar both spectral and spatial characteristics within a spatial window. If any individual feature point $\mathbf{x}_i^{spec}$ (or $\mathbf{x}_i^{Spat}$) is "far apart" from $\mathbf{x}_{im}^{Spec}$ (or $\mathbf{x}_{im}^{Spat}$), then $E_{im}^{Fus} = 0$. For example, suppose that the data point $\mathbf{x}_i^{Sta}$ belongs to a road and $\mathbf{x}_{im}^{Sta}$ belongs to a flat roof. Since, in practice, roads and roofs are often made with similar materials (e.g., asphalt), the corresponding data points are likely to have similar spectral characteristics ($E_{im}^{Spec} = 1$), but different spatial information (e.g., shape and size) ($E_{im}^{Spat} = 0$), so these two data points are not "close" (i.e., $E_{im}^{Fus} = 0$). Similarly, if $\mathbf{x}_i^{Sta}$ and $\mathbf{x}_{im}^{Sta}$ are taken from the grass areas and parking lot, respectively, they will have different spectral characteristics ($E_{im}^{Spec} = 0$), and even if they might be similar spatially ($E_{im}^{Spat} = 1$), the resulting $E_{im}^{Fus} = 0$ characterizing these two data points as "far apart". If the fusion graph was globally constructed by using the whole hyperspectral image or randomly selected samples like [76], one may find the kNN of a pixel belonging to a roof (e.g., shopping mall) $\mathbf{x}_i^{Fus}$ in pixels belonging to parking lots, because they have very similar spectral and spatial information even though they might not be spatially adjacent. By building a local fusion graph within a spatial window, the proposed LGF overcomes this limitation, and better models the local spatial neighborhood information. In addition, the proposed LGF is robust to image degradation (e.g., noise), which cannot be avoided during the hyperspectral image acquisition (especially when the spectral bands are in correspondence to windows in the electromagnetic spectrum in which the absorption of the atmosphere is high). The spectra of the same land cover type might exhibit a high variability. This is due to different factors such as the intrinsic variability of the reflectance, differences in illumination and image artifacts. However, typically the spectra of pixels belonging to the same object are correlated even if they might differ to those of objects of the same thematic class but located in other parts of the image for the above-mentioned reasons. Thus, by looking for the kNN within a spatial neighborhood can enforce to establish among pixels relations that are meaningful (in terms of representation of the objects). In a similar fashion, the approaches in [81, 83] showed better denoising results and efficient target detection by considering a local neighborhood.

Then, we can rearrange the edge of each stacked data point $\mathbf{E}_i^{Fus}$ into a sparse matrix $\mathbf{A}^{Fus}$ by using:

$$A_{ij}^{Fus} = \begin{cases} E_{ij}^{Fus}, & \text{if } j \in \mathcal{N}^{Fus}(i),\ j \in [1, \cdots, N] \\ 0, & \text{otherwise,} \end{cases} \tag{6.9}$$

The matrix $\mathbf{A}^{Fus} \in \mathbb{R}^{N \times N}$ represents the adjacency relation of all data points (e.g., full edge) built on the stacking features (i.e., $\mathbf{G}^{Fus} = (\mathbf{X}^{Sta}, \mathbf{A}^{Fus})$). We can obtain the transformation matrix to couple data fusion and dimension reduction of multisource data by solving the generalized eigenvalue problem similar as Eqs. (6.5) and (6.6).

The size of the sliding window has significant influence on the preservation of local spatial neighborhood information (e.g., texture). On the one hand, when the window size is too small, the neighborhood contains too few samples for properly modeling the local spatial information, and $\mathcal{N}_k^{Fus}(i)$ is composed of almost all data points within the window. On the other hand, if the window is too large, then the local spatial information might not be retrieved. In the case limit in which the neighborhood is the whole image, LGF equals to GDF (thus, GDF can be considered as a special case of LGF). In our experiments, we set the sliding window with a fixed intermediate size and change k nearest neighbors to obtain a satisfying result. By building a local fusion graph within a sliding window, we not only reduce both memory cost and computational complexity, but also increase the preservation of local spatial neighborhood information.

When dealing with high-resolution hyperspectral data, we can fast implement LGF by spatially downsampling the original HS image and EMP to the same ratio. The main spatial structure of the objects in a high-resolution remote sensing image will be preserved after spatially downsampling within a certain value of downsampling ratio. This way LGF will keep stable on the classification performances while reducing the computational complexity. The computational complexity can be reduced by a factor equal to the square of the spatial downsampling ratio. For example, if we downsample original HS image by a factor of R (e.g., $R = 4$) along both spatial directions, the total number of spatial pixels can be reduced to N/R^2, thus the computational complexity is reduced to $O(BNS^2/R^2)$.

6.3.2.1 Experimental Results and Analysis

The hyperspectral data of '*Pavia University*' is used to evaluate the performances of fusion approaches, as it has been widely used as a benchmark dataset for a classification task. The data was acquired by the ROSIS (Reflective Optics System Imaging Spectrometer) sensor over an urban areas in the city of Pavia, Italy, with 115 spectral bands in the wavelength range from 0.43 to 0.86μm and very fine spatial resolution of 1.3 meters by pixel. The image composed of 610 × 340 pixels contains 103 spectral channels after removal of noisy bands. This data set includes 9 land cover/use classes, see Fig. 6.11. Note that available training and testing set are given in Fig. 6.11 (# number of training samples /# number of test samples).

Prior to applying the morphological profiles to hyperspectral images, principal component analysis (PCA) was first applied to the original hyperspectral data set, and the first few principal components (PCs) (the first 3 PCs) were selected (representing 99% of the cumulative variance) to construct the EMP. A circular SE ranging from 1 to 10 with step size increment of 1 was used. 10 openings and closings were computed for each PC, resulting in an EMP of 63. We keep the same settings for SVM classifier as in Sect. 6.2.1.2 and compare the local graph fusion method LGF with the schemes of (1) Using original HS image (Raw); (2) Using EMP computed on the first 3 PCs of the original HS image (EMP); (3) Stacking all feature sources together, i.e., $\mathbf{X}^{Sta}$ (STA); (4) Stacking all the features extracted by PCA from each

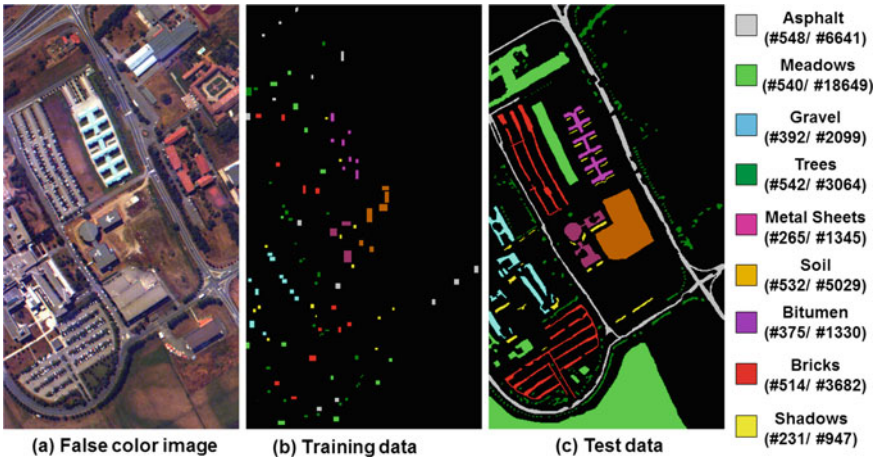


Fig. 6.11 Training and test data for *Pavia University* hyperspectral image

individual feature source (PCA); (5) Stacking all the features extracted by NWFE [77] from each individual feature source (NWFE), similar as [34]; (6) The GDF [76] with its extension to fuse two feature sources. The consumed time reported in our experiments includes both feature fusion and parameters optimization for SVM classification.

In order to make fair comparisons, for the approaches of PCA and NWFE in all our experiments, we use the best combination of the extracted spectral and the extracted spatial features for the classification. We search the best combination of the spectral and the spatial dimensions using the cross-validation according to the OA, with both the spectral dimension and the spatial dimension ranging from 2 to 40 (with step size increment of 2). The best combination is obtained when OA reaches the maximum. 5000 samples were randomly selected to build the global fusion graph in GDF, similar as we did in [76]. For the LGF, we first downsampled both original HS image and EMP of factor 5 on both spatial directions to speed up the processing time, and set the size of sliding window to 15×15. Table 6.5 reports the accuracies and consumed time as the downsampled size changes. The classification results using the best combination are shown in Table 6.4 and Fig. 6.12.

The results confirm that the integration of multisource features can improve the classification performance on HS images. Compared to the situation with single spectral or spatial feature source, the OA of stacking spectral and spatial features has 8.68–13.54% and 8.54–13.4% improvements for PCA and NWFE, respectively. The improvements of simply stacking original spectral and spatial features (STA) over only using the single spectral/spatial feature source are not significant, while increasing both the dimensionality and computational time. The LGF produced the better results, with OA improvements of 13.68–18.54% over only using the single spectral/spatial feature source, with OA improvements of 5–13.25% over stacking both the spectral and the spatial features by PCA, NWFE, and STA, and with 4.13% improvement over GDF.

Table 6.4 Classification accuracy for *Pavia University* hyperspectral data with SVM classifier. We built the local fusion graph of the proposed LGF on both the downsampled original HS image and EMP of factor 5

Number of Features	Raw 103	EMP 63	STA 166	PCA 42(2,40)	NWFE 40(8,32)	GDF 36	LGF 28
OA (%)	79.75	84.61	85.04	93.29	93.15	94.16	**98.29**
AA (%)	88.26	91.25	91.93	92.51	93.92	94.68	**98.66**
κ	0.747	0.802	0.809	91.05	0.909	0.923	**0.977**
Consumed Time (s)	65.05	27.28	75.42	25.67	876.78	22.69	**20.46**
Asphalt	84.17	95.04	93.14	**98.45**	97.74	93.89	**98.45**
Meadows	67.42	76.79	75.82	95.79	94.71	93.77	**97.95**
Gravel	73.70	82.18	80.51	77.99	82.28	77.94	**98.67**
Trees	94.78	96.77	97.10	84.86	**97.94**	93.24	93.77
Metal Sheets	99.63	**99.93**	**99.93**	99.85	**99.93**	99.78	99.78
Soil	92.30	71.72	82.04	80.25	73.61	95.67	**99.88**
Bitumen	91.20	99.77	99.77	**99.92**	99.85	99.25	**99.92**
Bricks	91.47	99.27	99.29	99.40	99.27	99.32	**99.51**
Shadows	99.68	99.79	99.79	96.09	**100**	99.26	**100**

From the class-specific accuracies, the EMP approach performed much better for most classes than the Raw approach, especially for the classes 'Asphalt' and 'Meadows,' with more than 10% improvement in accuracy. However, the EMP approach produced much worse accuracy in class 'Soil,' dropping by 20% compared to the Raw approach. By stacking spectral and spatial features extracted by PCA/NWFE, better accuracies were produced in class 'Meadows' (with improvements of 19–28.37% and 17.92–27.29%, respectively, over Raw and EMP); but the performance dropped significantly on classes 'Soil' compared to Raw. By building the fusion graph on randomly selected samples, the GDF approach consumes less time and performed much better than both Raw and EMP on classes 'Meadows' and 'Soil' (with OA improvements of 16.98–26.35% and 3.37–23.95%); but worse on class 'Gravel' compared to only using the spatial features. The LGF demonstrated better performance on almost all the classes than the methods that use single feature source (Raw and EMP), stacked multisource features (i.e., PCA, NWFE, and STA) and the GDF, and produced much better results on classes 'Gravel' and 'Soil'. For class 'Gravel', LGF had improvements of 16.49–24.97%, 16.39–20.68%, and 20.73% compared to the approaches using single feature source, stacked multisource features, and GDF, respectively.

The hyperspectral remote sensing data contains a wealth of spectral and spatial information. Only using single feature source is not enough for a reliable classification. When stacking multisource features extracted by methods like PCA and NWFE, it is not easy to select the optimal combination of the spectral and the spatial dimensions, as was also discussed by Fauvel et al. in [34]. These optimal combinations

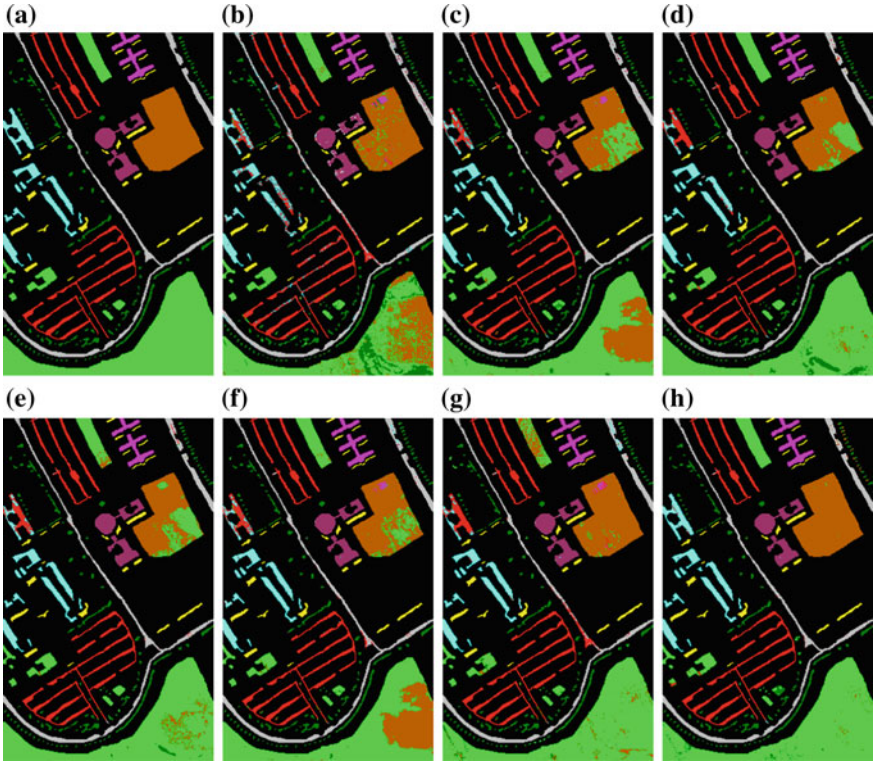


Fig. 6.12 Classification maps produced by different schemes. **a** Test samples, and thematic maps using **b** original HS data; **c** EMP of HS data; **d** PCA; **e** NWFE; **f** the stacked features STA; **g** GDF [76]; **h** the proposed LGF

of spectral and spatial dimensions are different for different data sets. Even for the same data set, when the training sample size changes, the combination of spectral and spatial dimensions will change.

Many approaches selected the optimal combination of spectral and the spatial dimensions according to the cumulative variance [34]. However, these approaches do not always work well. For example in PCA, the number of PCs which represent more than 99% of the cumulative variance depends on the statistical distribution of the data. The extracted PCs which represent 99% of the cumulative variance may not contain enough information of the data, resulting in a worse performance. When the data contain non-Gaussian noise, the number of PCs needed to reach 99% of the cumulative variance is higher, which may contain redundant information. Although some algorithms (e.g., cross-validation) can be used to find the best combination of dimensions, it increases the processing time. In our experiments, the elapsed time of searching the best combination is 7.99 and 7.63 h for PCA and NWFE, respectively.

Table 6.5 The accuracies and consumed time as the downsampled size of original feature sources increases. $DS_{3\times3}$ means we downsample both the original HS image and EMP of a factor 3 on both spatial directions

	$DS_{1\times1}$	$DS_{2\times2}$	$DS_{3\times3}$	$DS_{4\times4}$	$DS_{5\times5}$
OA (%)	97.11	98.23	97.56	97.63	98.29
AA (%)	97.04	98.58	98.12	98.11	98.66
κ	0.962	0.977	0.968	0.969	0.977
Consumed time (s)	1551.5	136.7	32.1	14.3	8.5

The performances of the LGF are less sensitive to the values of the free parameters. We keep the parameters (the number of nearest neighbors and the number of extracted features) the same for feature sources with different downsampling ratios, see Table 6.5. We get very similar classification results for *Pavia University*, with processing time dropping from 1551.5 to 8.5 s. Downsampling might cause a reduction of the intraclass heterogeneity (i.e., objects belonging to the same class will be more spectrally similar). If the training samples are taken far from the objects' edges, they will likely correspond to areas of an unique thematic class (i.e., they do not correspond to mixed pixels) leading to a simpler classification problem. Overall, local graph fusion technique effectively employs the local spatial information of different feature sources within a spatial window. This allows to obtain better performances in classification in particular with respect to the GDF approach which is global. In addition, with respect to this latter, we reduce both memory cost and computational complexity for graph building and increase robustness to image noise thanks to considering a small sliding window.

6.4 Discussion and Conclusions

We enter an era where multisource data is associated with high-impact commercial, social, biomedical and environmental datasets. Developing techniques for multisource data fusion can improve data quality, reduce information redundancy, extend the spatial and temporal coverages of the data, etc. Multisource data fusion has already benefited many disciplines, such as computer vision, medical imaging, sensor net works, robotics, intelligent system design, etc. This chapter focused on signal-level (e.g., pansharpening) and feature-level fusion of multisource remote sensing data. We introduced guided filter-based methods for hyperspectral image pansharpening, as well as exploited the use of graph-based fusion method to fuse multisource data to improve classification performances.

The objective of pansharpening is to fuse a Pan/RGB image with a multi-/hyperspectral one, to obtain an enhanced image with the high spatial resolution of the former and the high spectral resolution of the latter. The presented guided filter-based pansharpening method combines both component substitution and multiscale

decomposition, making a good balance on both spectral and spatial preservations. The presented pansharpening method is robust to image co-registration. By taking classification as an example to evaluate the quality of fused data, we found that fusion of multisource data enables improved classification performances by providing complementary information. However, developing effective methods for automatic pansharpening remains challenging, due to the fast development of sensor technologies, e.g., extremely high-resolution optical, SAR, and LiDAR sensors from airborne or UAVs platforms. Precise co-registration and new pansharpening techniques are required to improve the spatial resolution, as well as retain the spectral fidelity of original hyperspectral data, during pixel-level fusion.

Feature-level fusion aims at optimizing the complementary information from multisource data to achieve better decision than using single data source. Feature-level fusion requires feature extraction from different sources. Taking land cover/use classification as an example, the most straightforward way to perform this fusion may be to concentrate all data source together and use this concentrated data as the input of a classifier. However, such fusion methods do not take into account the differences between feature sources and may lead to problems like the curse of dimensionality and excessive computation time. The presented graph-based fusion methods coupled dimension reduction and data fusion together for multisource data. Our graph-based methods combined multiple feature sources through a fused graph, which explains the relations between data points in different data sources and can be seen as a way to model the embedding in the manifold in which the data lie. The graph-based feature-level fusion methods proved to overcome the conventional approaches of stacking different feature sources together in terms of classification accuracy. However, developing effective methods for automatic fusion and interpretation of the multisource data is still very difficult. 'Big data' from remote sensing requires new techniques on both feature extraction and fusion which are in computation efficiency and effectiveness.

The discussion on remote sensing data fusion will continue in the next chapter, where the focus will be on mathematical models for supervised classification of multisensor, multiscale, and multiresolution imagery.

Acknowledgements The authors would like to thank Telops Inc. (Québec, Canada), the Hyperspectral Image Analysis group and the NSF Funded Center for Airborne Laser Mapping (NCALM) at the University of Houston, and Prof. Paolo Gamba from the University of Pavia (Italy) for providing the data sets used in this Chapter. This work was supported by the Fund for Scientific Research in Flanders (FWO) project G037115N "Data Fusion for Image Analysis in Remote Sensing." Wenzhi Liao is a postdoctoral fellow of the Research Foundation Flanders (FWO-Vlaanderen) and acknowledges its support.

References

1. Hall, L., Llinas, J.: An introduction to multisensor data fusion. Proc. IEEE **85**, 6–23 (1997)
2. Pohl, C., Van Genderen, J.L.: Multisensor image fusion in remote sensing: concepts, methods and applications. Int. J. Remote Sens. **19**, 823–854 (1998)

3. Zhang, J.: Multi-source remote sensing data fusion: status and trends. Int. J. Image Data Fus. **1**(1), 5–24 (2010)
4. Gamba, P.: Image and data fusion in remote sensing of urban areas: status issues and research trends. Int. J. Image Data Fus. **5**(1), 2–12 (2014)
5. Dalla Mura, M., Prasad, S., Pacifici, F., Gamba, P., Chanussot, J., Benediktsson, J.: Challenges and opportunities of multimodality and data fusion in remote sensing. Proc. IEEE **103**(9), 1585–1601 (2015)
6. Lahat, D., Adali, T., Jutten, C.: Multimodal data fusion: an overview of methods, challenges and prospects. Proc. IEEE **103**(9), 1449–1477 (2015)
7. Gomez-Chova, L., Tuia, D., Moser, G., Camps-Valls, G.: Multimodal classification of remote sensing images: a review and future directions. Proc. IEEE **103**(9), 1560–1584 (2015)
8. Luo, R.C., Kay, M.G.: A tutorial on multisensor integration and fusion. In:16 Annual Conference of IEEE Industrial Electronics Society, 1990, pp. 707–722 (1990)
9. Loncan, L., Almeida, L.B., Bioucas Dias, J., et al.: Hyperspectral pansharpening: a review. IEEE Geosci. Remote Sens. Mag. **3**(3), 27–46 (2015)
10. Carper, W., Lillesand, T.M., Kiefer, P.W.: The use of intensity- hue-saturation transformations for merging SPOT panchromatic and multispectral image data. Photogramm. Eng. Remote Sens. **56**(4), 459–467 (1990)
11. Tu, T.M., Su, S.C., Shyu, H.C., Huang, P.S.: A new look at IHS-like image fusion methods. Inf. Fus. **2**(3), 117–186 (2001)
12. Shettigara, V.: A generalized component substitution technique for spatial enhancement of multispectral images using a higher resolution data set. Photogramm. Eng. Remote Sens. **58**(5), 561–567 (1992)
13. Shah, V., Younan, N., King, R.: An efficient pan-sharpening method via a combined adaptive PCA approach and contourlets. IEEE Trans. Geosci. Remote Sens. **46**(5), 1323–1335 (2008)
14. Mallat, S.: A theory for multiresolution signal decomposition: the wavelet representation. IEEE Trans. Pattern Anal. Mach. Intell. **11**(7), 674–693 (1989)
15. Nason, G.P., Silverman, B.W.: The stationary wavelet transform and some statistical applications in Wavelets and Statistics. In: A. Antoniadis, G.Oppenheim (ed.), vol. 103, pp. 281–299 Springer, New York (1995)
16. Burt, P.J., Adelson, E.H.: The Laplacian pyramid as a compact image code. IEEE Trans. Commun. **31**(4), 532–540 (1983)
17. Do, M.N., Vetterli, M.: The contourlet transform: an efficient directional multiresolution image representation. IEEE Trans. Image Process. **14**(12), 2091–2106 (2005)
18. Starck, J., Murtagh, F.: The undecimated wavelet decomposition and its reconstruction. IEEE Trans. Image Process. **16**(2), 297–309 (2007)
19. Ballester, C., Caselles, V., Igual, L., et al.: A variational model for P+XS image fusion. Int. J. Comput. Vis. **5969**(1), 43–58 (2006)
20. Palsson, F., Sveinsson, J., Ulfarsson, M., Benediktsson, J.: A new pansharpening algorithm based on total variation. IEEE Geosci. Remote Sens. Lett. **11**(1), 318–322 (2014)
21. He, X., Condat, L., Bioucas Dias, J., et al.: A new pansharpening method based on spatial and spectral sparsity priors. IEEE Trans. Image Process. **23**(9), 4160–4174 (2014)
22. Moeller, M., Wittman, T., Bertozzi, A.: A variational approach to hyperspectral image fusion. In: SPIE Defense, Security, and Sensing (2009)
23. Vivone, G., Alparone, L., Chanussot, J., et al.: Multi-resolution analysis and component substitution techniques for hyperspectral pansharpening. In: 2014 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), 2649–2652 (2014)
24. Yokoya, N., Yairi, T., Iwasaki, A.: Coupled nonnegative matrix factorization unmixing for hyperspectral and multispectral data fusion. IEEE Trans. Geosci. Remote Sens. **50**(2), 528–537 (2012)
25. Wei, Q., Dobigeon, N., Tourneret, J.Y.: Bayesian fusion of multi- band images. IEEE J. Select. Top. Signal Process. **9**(6), 1117–1127 (2015)
26. Simoes, M., Bioucas-Dias, J., Almeida, L., Chanussot, J.: A convex formulation for hyperspectral image superresolution via subspace-based regularization. IEEE Trans. Geosci. Remote Sens. **53**(6), 3373–3387 (2015)

27. Wei, Q., Bioucas-Dias, J., Dobigeon, N., Tourneret, J.Y.: Hyperspectral and multispectral image fusion based on a sparse representation. IEEE Trans. Geosci. Remote Sens. **53**(7), 3658–3668 (2015)
28. Liao, W., Huang, X., Coillie, F., et al.: Processing of Multiresolution Thermal Hyperspectral and Digital Color Data: Outcome of the 2014 IEEE GRSS Data Fusion Contest. IEEE J. Select. Top. Appl. Earth Observ. Remote Sens. **8**(6), 2984–2996 (2015)
29. Liao, W., Huang, X., Coillie, F., Guy, T., Scheunders, P., Pizurica, A., Philips, W. Two-stage fusion of thermal hyperspectral and visible RGB image by PCA and Guided filter. In: 7th workshop on hyperspectral image and signal processing: evolution in remote sensing (WHISPERS 2015), Tokyo, Japan (2015)
30. Zhu, X., Grohnfeldt, C., Bamler, R.: Exploiting joint sparsity for pan-sharpening: the J-sparse FI algorithm. IEEE Trans. Geosci. Remote Sens. **54**(5), 2664–2681 (2016)
31. He, K., Sun, J., Tang, X.: Guided image filtering. IEEE Trans. Pattern Anal. Mach. Intell. **35**(6), 1397–1409 (2013)
32. Burges, C.J.C.: A tutorial on support vector machines for pattern recognition. Data Min. Knowl. Discov. **2**, 121–167 (1998)
33. Chang, C.C., Lin, C.J.: (2001). LIBSVM: A Library for Support Vector Machines. ACM Trans. Intell. Syst. Technol., 2(3), 27:1–27:27, 2011. http://www.csie.ntu.edu.tw/~cjlin/libsvm
34. Fauvel, M., Benediktsson, J.A., Chanussot, J., Sveinsson, J.R.: Spectral and spatial classification of hyperspectral data using SVMs and morphological profile. IEEE Trans. Geosci. Remote Sens. **46**(11), 3804–3814 (2008)
35. Swatantrana, A., Dubayaha, R., Robertsb, D., Hoftona, M., Blairc, J.B.: Mapping biomass and stress in the Sierra Nevada using lidar and hyperspectral data fusion. Remote Sens. Environ. **115**(11), 2917–2930 (2011)
36. Koetz, B., Sun, G., Morsdorf, F., Ranson, K.J., et al.: Fusion of imaging spectrometer and LIDAR data over combined radiative transfer models for forest canopy characterization. Remote Sens. Environ. **106**(4), 449–459 (2007)
37. Dalponte, M., Bruzzone, L., Gianelle, D.: Fusion of Hyperspectral and LIDAR Remote Sensing Data for Classification of Complex Forest Areas. IEEE Trans. Geosci. Remote Sens. **46**(5), 1416–1427 (2008)
38. Naidooa, L., Choa, M.A., Mathieua, R., Asnerb, G.: Classification of savanna tree species, in the Greater Kruger National Park region, by integrating hyperspectral and LiDAR data in a Random Forest data mining environment. ISPRS J. Photogramm. Remote Sens. **69**, 167–179 (2012)
39. Pedergnana, M., Reddy Marpu, P., Dalla Mura, M., Benediktsson, J.A., Bruzzone, L.: Classification of remote sensing optical and LiDAR data using extended attribute profiles. IEEE J. Select. Top. Signal Process. **6**(7), 856–865 (2012)
40. Khodadadzadeh, M., Li, J., Prasad, M., Plaza, A.: Fusion of Hyperspectral and LiDAR Remote Sensing Data Using Multiple Feature Learning. IEEE J. Select. Top. Appl. Earth Observ. Remote Sens. **8**(6), 2971–2983 (2015)
41. Zhang, Y., Prasad, S.: Multisource geospatial data fusion via local joint sparse representation. IEEE Trans. Geosci. Remote Sens. **54**(6), 3265–3276 (2016)
42. Liao, W., Bellens, R., Pizurica, A., Gautama, S., Philips, W.: Generalized graph-based fusion of hyperspectral and LiDAR data using morphological features. IEEE Geosci. Remote Sens. Lett. **12**(3), 552–556 (2015)
43. Camps-Valls, G., Gomez-Chova, L., Munoz-Mari, J., Vila-Frances, J., Calpe-Maravilla, J.: Composite kernels for hyperspectral image classification. IEEE Geosci. Remote Sens. Lett. **3**(1), 93–97 (2006)
44. Fauvel, M., Chanussot, J., Benediktsson, J.: A spatial-spectral kernel-based approach for the classification of remote-sensing images. Pattern Recognit. **45**(1), 381–392 (2012)
45. Li, J., Marpu, P.R., Plaza, A., Bioucas-Dias, J.M., Benediktsson, J.: Generalized composite Kernel framework for hyperspectral image classification. IEEE Trans. Geosci. Remote Sens. **51**(9), 4816–4829 (2013)

46. Voisin, A., Krylov, V.A., Moser, G., Serpico, S.B., Zerubia, J.: Supervised Classification of Multisensor and Multiresolution Remote Sensing Images with A Hierarchical Copula-based Approach. IEEE Trans. Geosci. Remote Sens. **52**(6), 3346–3358 (2014)
47. Tuia, D., Volpi, M., Trolliet, M., Camps-Valls, G.: Semisupervised manifold alignment of multimodal remote sensing images. IEEE Trans. Geosci. Remote Sens. **52**(12), 7708–7720 (2014)
48. Fang, L., Li, S., Kang, X., Benediktsson, J.: Spectral-spatial hyperspectral image classification via multiscale adaptive sparse representation. IEEE Trans. Geosci. Remote Sens. **52**(12), 7738–7749 (2014)
49. Gunatilaka, A.H., Baertlein, B.A.: Feature-level and decision-level fusion of noncoincidently sampled sensors for land mine detection. IEEE Trans. Pattern Anal. Mach. Intell. **23**(6), 577–589 (2001)
50. Calhoun, V.D., Adali, T., Pearlson, G.D., Kiehl, K.A.: Neuronal chronometry of target detection: fusion of hemodynamic and event-related potential data. NeuroImage **30**(2), 544–553 (2006)
51. Calhoun, V.D., Adali, T., Liu, J.: A feature-based approach to combine functional MRI, structural MRI, and EEG brain imaging data. In: 2006 Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS), New York (2006)
52. Correa, N.M., Li, Y.O., Adali, T., Calhoun, V.D.: Canonical correlation analysis for feature-based fusion of biomedical imaging modalities and its application to detection of associative networks in schizophrenia. IEEE J. Select. Top. Signal Process. **2**(6), 998–1007 (2008)
53. Jagadeesan, A., Thillaikkarasi, T., Duraiswamy, K.: Protected bio-cryptography key invention from multimodal modalities: feature level fusion of fingerprint and Iris. Eur. J. Sci. Res. **49**(4), 484–502 (2011)
54. Conti, V., Militello, C., Sorbello, F., Vitabile, S.: A frequency-based approach for features fusion in fingerprint and iris multi-modal biometric identification systems. IEEE Trans. Syst. Man Cybern. C **40**(4), 384–395 (2010)
55. Nagar, A., Nandakumar, K., Jain, A.K.: Multibiometric cryptosystems based on feature-level fusion. IEEE Trans. Inf. Forensics Secur. **7**(1), 255–268 (2012)
56. Camps-Valls, G., Gomez-Chova, L., et al.: Kernel-based framework for multitemporal and multisource remote sensing data classification and change detection. IEEE Trans. Geosci. Remote Sens. **46**(6), 1822–1835 (2008)
57. Tuia, D., Ratle, F., Pozdnoukhov, A., Camps-Valls, G.: Multi-source composite kernels for urban image classification. IEEE Geosci. Remote Sens. Lett. **7**(1), 88–92 (2010)
58. Tuia, D., Camps-Valls, G., Matasci, G., Kanevski, M.: Learning relevant image features with multiple kernel classification. IEEE Trans. Geosci. Remote Sens. **48**(10), 3780–3791 (2010)
59. Gomez-Chova, L., Camps-Valls, G., Bruzzone, L., Calpe-Maravilla, J.: Mean map kernel methods for semisupervised cloud classification. IEEE Trans. Geosci. Remote Sens. **48**(1), 207–220 (2010)
60. Volpi, M., Camps-Valls, G., Tuia, D.: Spectral alignment of cross-sensor images with automated kernel canonical correlation analysis. ISPRS J. Photogramm. Remote Sens. **107**, 50–63 (2015)
61. Dalla Mura, M., Benediktsson, J.A., Waske, B., Bruzzone, L.: Extended profiles with morphological attribute filters for the analysis of hyperspectral data. Int. J. Remote Sens. **31**(22), 5975–5991 (2010)
62. Dalla Mura, M., Villa, A., Benediktsson, J.A., Chanussot, J., Bruzzone, L.: Classification of hyperspectral images by using extended morphological attribute profiles and independent component analysis. IEEE Geosci. Remote Sens. Lett. **8**(3), 541–545 (2011)
63. Liao, W., Dalla Mura, M., Chanussot, J., Pizurica, A.: Fusion of Spectral and Spatial Information for Classification of Hyperspectral Remote Sensed Imagery by Local Graph. IEEE J. Select. Top. Appl. Earth Observ. Remote Sens. **9**(2), 583–594 (2016)
64. Blaschke, T.: Object based Image Analysis for Remote Sensing. ISPRS J. Photogramm. Remote Sens. **65**(1), 2–16 (2010)
65. Soille, P.: Morphological Image Analysis, Principles and Applications, 2nd edn. Springer, Berlin (2003)

66. Benediktsson, J., Palmason, J., Sveinsson, J.R.: Classification of hyperspectral data from urban areas based on extended morphological profiles. IEEE Trans. Geosci. Remote Sens. **43**(3), 480–491 (2005)
67. Dalla Mura, M., Benediktsson, J., Waske, B., Bruzzone, L.: Morphological attribute profiles for the analysis of very high resolution images. IEEE Trans. Geosci. Remote Sens. **48**(10), 3747–3762 (2010)
68. Huang, X., Liu, H., Zhang, L.: Spatiotemporal detection and analysis of urban villages in mega city regions of china using high-resolution remotely sensed imagery. IEEE Trans. Geosci. Remote Sens. **53**(7), 3639–3657 (2015)
69. Bruzzone, L., Bovolo, F.: A novel framework for the design of change-detection systems for very-high-resolution remote sensing images. Proc. IEEE **101**(3), 609–630 (2013)
70. Braun, A.C., Rojas, C., et al.: Design of a Spectral-Spatial Pattern Recognition Framework for Risk Assessments Using Landsat Data-A Case Study in Chile. IEEE J. Select. Top. Appl. Earth Observ. Remote Sens. **7**(3), 917–928 (2014)
71. Liao, W., Bellens, R., Pižurica, A., Philips, W., Pi, Y.: Classification of Hyperspectral Data Over Urban Areas Using Directional Morphological Profiles and Semi-Supervised Feature Extraction. IEEE J. Select. Top. Appl. Earth Observ. Remote Sens. **5**(4), 1177–1190 (2012)
72. Bellens, R., Gautama, S., Martinez-Fonte, L., Philips, W., Chan, J.C.-W., Canters, F.: Improved classification of VHR images of urban areas using directional morphological profiles. IEEE Trans. Geosci. Remote Sens. **46**(10), 2803–2812 (2008)
73. Scholkopf, B., Smola, A.J., Muller, K.R.: Nonlinear component analysis as a kernel eigenvalue problem. Neural Comput. **10**, 1299–1319 (1998)
74. Belkin, M., Niyogi, P.: Laplacia Eigenmaps and Spectral Techniques for Embedding and Clustering. In: Advances in Neural Information Processing Systems 14, 585–591, MIT Press, British Columbia, Canada (2002)
75. He, X.F., Niyogi, P.: Locality preserving projections. In: Advances in Neural Information Processing Systems 16, pp. 153–160. MIT Press, Cambridge (2004)
76. Debes, C., Merentitis, A., Heremans, R., et al.: Hyperspectral and LiDAR Data Fusion: Outcome of the 2013 GRSS Data Fusion Contest. IEEE J. Select. Top. Appl. Earth Observ. Remote Sens. **7**(6), 2405–2418 (2014)
77. Kuo, B.C., Landgrebe, D.A.: Nonparametric weighted feature extraction for classification. IEEE Trans. Geosci. Remote Sens. **42**(5), 1096–1105 (2004)
78. Tarabalka, Y., Benediktsson, J.A., Chanussot, J.: Spectral-spatial classification of hyperspectral imagery based on partitional clustering techniques. IEEE Trans. Geosci. Remote Sens. **47**(8), 2973–2987 (2009)
79. Li, J., Bioucas-Dias, J.M., Plaza, A.: Spectral-spatial hyperspectral image segmentation using subspace multinomial logistic regression and Markov random field. IEEE Trans. Geosci. Remote Sens. **50**(3), 809–823 (2012)
80. Camps-Valls, G., Shervashidze, N., Borgwardt, K.M.: Spatio-spectral remote sensing image classification with graph Kernels. IEEE Geosci. Remote Sens. Lett. **7**(4), 741–745 (2010)
81. Chen, G., Qian, S.E.: Dimensionality reduction of hyperspectral imagery using improved locally linear embedding. J. Appl. Remote Sens. **1**, 1–10 (2007)
82. Jimenez, M.D., Prelcic, N.: Linear boundary extensions for einite length signals and paraunitary two-channel filterbanks. IEEE Trans. Signal Process **52**(11), 3213–3226 (2004)
83. Chen, G., Bui, T.D., Krzyzak, A.: Image denoising with neighbour dependency and customized wavelet and threshold. Pattern Recogn. **38**(1), 115–124 (2005)

Chapter 7
Remote Sensing Data Fusion: Markov Models and Mathematical Morphology for Multisensor, Multiresolution, and Multiscale Image Classification

Jon A. Benediktsson, Gabriele Cavallaro, Nicola Falco, Ihsen Hedhli, Vladimir A. Krylov, Gabriele Moser, Sebastiano B. Serpico and Josiane Zerubia

Abstract Current and forthcoming sensor technologies and space missions are providing remote sensing scientists and practitioners with an increasing wealth and variety of data modalities. They encompass multisensor, multiresolution, multiscale, multitemporal, multipolarization, and multifrequency imagery. While they represent remarkable opportunities for the applications, they pose important challenges to the development of mathematical methods aimed at fusing the information conveyed by the input multisource data. In this framework, the present chapter continues the discussion of remote sensing data fusion, which was opened in the previous chapter. Here, the focus is on data fusion for image classification purposes. Both methodological issues of feature extraction and supervised classification are addressed. On both respects, the focus is on hierarchical image models rooted in graph theory. First, multilevel feature extraction is addressed through the latest advances in Mathematical Morphology and attribute profile theory with respect to component trees and trees of

J.A. Benediktsson (✉) · G. Cavallaro
University of Iceland, Saemundargotu 2 - 101, Reykjavik, Iceland
e-mail: benedikt@hi.is

N. Falco
Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, CA 94720, USA
e-mail: nicolafalco@lbl.gov

I. Hedhli · V.A. Krylov · G. Moser · S.B. Serpico
University of Genoa, Via Opera Pia 11a, 16145 Genoa, Italy
e-mail: gabriele.moser@unige.it

I. Hedhli · V.A. Krylov · J. Zerubia
Université Côte d'Azur, Inria, BP 93, 2004, Route des Lucioles, 06902 Sophia Antipolis Cedex, France
e-mail: josiane.zerubia@inria.fr

G. Moser and J. Zerubia (eds.), *Mathematical Models for Remote Sensing Image Processing*, Signals and Communication Technology,
https://doi.org/10.1007/978-3-319-66330-2_7

shapes. Then, joint supervised classification of multisensor, multiscale, multiresolution, and multitemporal imagery is formulated through hierarchical Markov random fields on quad-trees. Examples of experimental results with data from current VHR optical and SAR missions are shown and analysed.

7.1 Multisource Data Fusion for Image Classification

As pointed out in the previous chapters, current remote sensing systems allow a remarkable variety of data typologies to be collected. These data differ in sensor technology (passive: optical and thermal; and active: radar and laser), spatial resolution (from a few kilometers with weather satellites to below a meter with satellite VHR missions – see Chaps. 3 and 4 – or a few centimeters with drones and other aerial platforms – see Chaps. 5 and 6), and spectral resolution (from one panchromatic channel to multispectral or hyperspectral channels – see Chaps. 2 and 6) [76]. Jointly exploiting these data represents a crucial and challenging opportunity. Crucial because data from different sources often convey complementary information, therefore a proper joint use may allow more complete and accurate results to be obtained than using the single sources separately. Challenging because multisource data (also named multimodal data) are often heterogeneous, therefore advanced mathematical methods for modelling, learning, and merging are necessary to take benefit from them.

The fusion of multisensor and multiresolution remote sensing data has been addressed in the previous chapter with a focus on signal-level techniques that generate enhanced image products [83, 86]. In the present chapter, the discussion is continued by addressing how to exploit images associated with different sensors and/or spatial scales to generate classification products.

As mentioned in several of the previous chapters, land cover or land use classification methodologies play a major role in remote sensing. Thorough discussions of classification concepts can be found in pattern recognition textbooks [14], and the basics have also been recalled in Chap. 2. Here, we shall assume that the reader is familiar with these concepts and focus on the case of the classification of multisource remote sensing images. Consistently with the usual pattern-recognition pipeline, we shall discuss both (i) *classification* techniques that operate with multisensor or multiscale features as inputs, and (ii) *feature extraction* methods that derive multiscale descriptors from the input imagery. The next subsections will review the main literature approaches related to (i) and (ii). Then, examples of advanced mathematical methods for multilevel feature extraction through morphological operators, and then, multisensor, multiresolution, and multitemporal classification through hierarchical Markov models will be discussed in Sects. 7.2 and 7.3, respectively. Similar to the approaches described in Chap. 3 with regard to VHR optical image analysis, both families of methods substantially rely on graph-theoretic hierarchical image representations through suitable tree topologies. Further, more focused, analysis on the specific case of multitemporal data can also be found in Chaps. 8 and 9.

7.1.1 Multiscale Feature Extraction

On one hand, in high resolution images, the high geometrical detail of the captured scenes results in a substantial increase in information, allowing for detailed analysis of complex structures that characterize the scene under investigation. On the other hand, the improved spatial resolution adds a certain spectral variability to the pixels that belong to the same object or class. Therefore, approaches that exploit the spectral information at single pixel level are poorly effective for the characterization of scenes in which heterogeneous structures are composed by several pixels, while strategies that include contextual information represented by pixels and their spatial neighbourhood systems become necessary.

In the last decade, several advances have been made in exploring multiscale strategies based on image processing to describe patterns such as shape and texture at different levels of abstraction. Scale is indeed an important aspect since it directly contributes to the effectiveness in modelling the spatial context [93]. However, when in presence of heterogeneous structures, the identification of an optimal scale is far from being trivial [4]. However, a multiscale strategy allows the pixel neighbourhood system to be defined adaptively. A simple and effective strategy for including spatial information is to enlarge the feature space, which is used as input to a learning algorithm, by adding features that provide information on the spatial arrangement [25].

Spatial features can be computed by exploiting image segmentation and object-based image analysis [15, 16] procedures by partitioning an image into non-overlapping regions/objects according to a homogeneity criterion. Several segmentation procedures have been introduced in the literature, including hierarchical and multiscale clustering strategies [23, 79]. In a hierarchical segmentation, a region of interest can be represented by multiple sub-regions (segments) in finer levels of detail and merged to the surrounding regions at coarser levels of detail, where the merging of segments is usually performed based on similarity measures. Optimized approaches based on the integration of spatial and spectral information through spectral clustering are also available in the literature [140], and recently exploited in hyperspectral image classification [138, 157].

A different approach to obtain spatial feature is through filtering. In the literature, a wide range of filters can be found [34]. In particular, filters based on statistical measures [141], which focus on the spatial distribution of grey values, have been extensively exploited for texture feature extraction. Statistical measures refer to first-order statistics (i.e. average intensity) [141] and second-order statistics, including the grey-level co-occurrence matrix (GLCM) [62]. By varying the window parameters, such as direction and size, GLCM was effectively exploited for the extraction of multiscale features for urban characterization in SAR image [45] and high resolution panchromatic images [110].

Another category of filters is represented by those based on signal processing, such as wavelet transforms [91, 122]. Wavelet decompositions produce a multiresolution representation of the original scene and have been successfully exploited for

numerous image processing and analysis tasks, including SAR image classification [55, 132] and optical image restoration [67].

A different typology of filters is represented by those defined as structural filters, such as morphological filters. They have been proposed within the Mathematical Morphology framework based on non-linear operators [130] and are an extension of Minkowski's set theory [130]. Erosion and dilation are considered the basic morphological operators, which are based on a moving window (or kernel), called structuring element (SE). In general, dilation causes objects to dilate or grow in size, whereas erosion causes objects to shrink. The effect of the filtering, i.e., the way objects dilate or shrink, depends upon the choice of the SE, i.e., shape and size. By combining dilation and erosion we obtain the closing and opening operators. Those operators are used to remove objects that cannot contain the SE, while preserving objects with a similar shape as the SE. However, this filtering process also yields distortions and may degrade the geometrical characteristics of the objects. Operators by reconstruction, which are based on geodesic transformations, were introduced to prevent this issue, allowing the preservation of the geometrical properties of those regions not affected by the filtering. By defining a family of increasing values corresponding to the size of the SE, it is possible to obtain a vector (morphological profile) of filtered images that provides a multiscale decomposition of the scene under investigation. Morphological profiles have been exploited for multiscale analysis in different applications, such as image segmentation, object extraction [1, 116], VHR image classification [10, 144], as well as hyperspectral image analysis and classification [9, 117, 148].

More recently, attribute filters have been introduced as an extension of morphological filters [20]. Attribute filters are connected operators, whose filtering is based on the evaluation of an attribute (i.e., a measure) computed on connected regions [133]. The filtering results in a simplification of the image, in which regions that do not fulfill a given predicate are merged to the surrounding regions. In contrast to SE-based filters, filters based on attributes allow the extraction of complementary contextual information according to the chosen attribute, and are more flexible in terms of spatial modelling. As for morphological filters, a multiscale image representation can be obtained by sequential application of attribute filters according to a pre-defined set of thresholds [30, 40]. Such structures can be further extended to multi-channel images [29, 39]. In this case, attribute profiles are computed for all the available channels or a subset of them. Due to their aforementioned properties, these operators have proven their efficiency in addressing crucial remote sensing tasks, ranging from image classification [41, 50, 120] to change detection [51]. Furthermore, in addition to their multiscale fusion capabilities, morphological and attribute filters have also proven successful for feature-level fusion (e.g., in combination with the local graph model described in Chap. 6), multisensor fusion [113, 139], and multitemporal fusion [33].

More information on the effectiveness of these operators can also be found in the recent literature, where comparison between different approaches for spatial feature extraction are provided [25, 53, 81].

7.1.2 Multisensor and Multiresolution Image Classification

The complementary properties of panchromatic and multispectral data have been discussed in depth in the previous chapter. Another major example of complementarity among remote sensing data sources is represented by passive and active sensors. Through passive sensors, the incoming spectral radiance, indirectly related to the reflectance, emittance, and temperature of the observed surface, is measured in several bands of the visible and infrared wavelength ranges (see Chaps. 2 and 3). Noise variance in the resulting data is small in the case of recent sensors, and visual photointerpretation is relatively easy, but acquisition capability is strongly affected by Sun illumination and cloud cover. An active SAR sensor transmits a microwave pulse toward the target area and collects the backscattered return (see Chaps. 4 and 5). The resulting data are determined by the scattering mechanism on the imaged area and influenced by roughness, soil moisture, and presence of strong scatterers. Day-and-night acquisition is feasible with little to no impact of cloud cover but photointerpretation is substantially more difficult than with optical images, and automated analysis is made complicated by speckle. Thus, the properties of these two data sources are intrinsically complementary. Active LiDAR sensors transmit laser pulses and again receive the backscattered signals (see Chap. 1). Terrain elevation features are typically extracted from LiDAR returns, thus providing a further complementary source of information with respect to 2D data. In this book, the focus is on 2D remote sensing image analysis, so here we shall not discuss the fusion of 2D and 3D data and we shall address the fusion of multimodal 2D imagery, specifically with the primary purpose of supervised classification.

Joint classification of multisource data has a long tradition in remote sensing, and many mathematical models and methods, including evidential, statistical, kernel, neural, decision fusion, and Markovian approaches, have been proposed [61].

First, when a probabilistic Bayesian formulation [14] is used for classification, a major difficulty in the case of multisource data is that, while accurate parametric models are known for the marginal statistics of the data originating from the individual sources (e.g., multivariate Gaussian for optical data [76] or Gamma and more sophisticated models for SAR data [73]), parametric models for their joint statistics are usually unavailable, except in special cases (e.g., [85]). A simple workaround might be to assume that the features associated with distinct sources are independent when conditioned to each class, but with this choice, mutual dependencies are neglected [112]. Alternately, advanced statistical tools are used to combine estimates of the marginal distributions of single features or subsets of features with models of their dependence. These include the theory of copula functions [72], dependence trees [43], or meta-Gaussian distributions [136].

In contrast to parametrically modeling the class-conditional statistics, fully non-parametric classifiers [14] have also been widely used for multisource image classification. After the features extracted from all data sources are collected in a unique vector, the resulting "stacked vector" can be fed as input to non-parametric classifiers because they are applicable to data with arbitrary joint statistics. Many formulations

have been proposed in this framework, involving multilayer perceptron neural networks [99], adaptive resonance theory [57], neurofuzzy architectures [3], support vector machines [42], k nearest neighbor classifiers [111], Parzen density estimators [43], logistic regression [158], and composite kernel functions with associated Hilbert spaces (see also Chap. 10) [24]. These methods have been used to classify in stacked feature spaces originating from optical, SAR, hyperspectral, and LiDAR data. More recently, deep learning and convolutional neural networks [78], which have been increasingly popular in remote sensing lately, have also been applied to classification problems involving multiscale [89], multifeature [94], and multisensor [142] fusion.

A further approach is Dempster-Shafer's mathematical theory of evidence, which quantifies the contributions of the individual sources in terms of appropriate mass of evidence, belief, and plausibility functions, and introduces a set of algebraic rules to combine these quantities into the output inference result [18]. Evidential methods have been proposed for the classification of multisensor optical-LiDAR data [123], multisensor optical-SAR images [84], multiresolution optical imagery [77], as well as for the fusion of remote sensing and ancillary data [121].

Whereas the previous approaches provide mathematical formulations for a unique classifier with multiple input sources, the decision fusion approach is based on the idea of first separately applying distinct classifiers to individual sources and then combining the classifier outputs. This multiple-classifier or classifier-ensemble approach moves the complexity of the fusion process to the output combination rule. Mathematical models for this rule have been proposed on the basis of voting schemes [100], consensus theory [7], Bayesian criteria [13], kernel functions [31], neural networks [49], fuzzy logic [52], graph theory [5], and attractor dynamics [17]. They have been applied to multisensor data sets including multispectral, hyperspectral, SAR, and LiDAR sources. Applications to multiscale optical imagery can also be found [48]. Hybrid approaches that integrate decision fusion and kernel learning have also been developed [156]. Random forest [21], which is a currently popular classifier because of its robustness to overfitting and low computational burden, and which has been applied to multisensor image classification as well [150], is a decision fusion method in itself because it combines the outputs of a random ensemble of tree classifiers. Extensions of random forest and alternate tree ensembles can be found in [100, 153].

When high spatial resolution is involved, multisource fusion methods based on probabilistic graphical models can be used as an alternative to the approach of contextual feature extraction described in the previous section. MRF models, which have already been discussed in Chap. 4, play a prominent role in this regard [82]. The opportunity to incorporate multiple terms, each associated with one input information source, into the energy function of an MRF model, makes MRFs powerful data fusion tools [129]. This approach has been applied to the fusion of both multisensor [60] and multiscale [103] data. Furthermore, the opportunity to define MRFs on hierarchical graphs, such as quad-trees [75], binary partition trees [118], or more irregular topologies [128], also makes Markovian models into natural multiscale and multiresolution fusion tools. Indeed, recent approaches based on hierarchical MRFs

have been formulated for classifying data collected at multiple input spatial resolutions at the same time [102], in a multitemporal series [64], or even by different optical/SAR sensors [63, 65].

An extension of MRFs is represented by conditional random fields (CRFs) [137], which postulate a Markovian formulation for the global posterior distribution directly, with the aim of gaining additional flexibility as compared to MRFs. Pixelwise (unary), pairwise (binary), and possibly higher order pixel dependencies are usually characterized through case-specific parametric models. In [66], a CRF is proposed for the joint multiscale and multitemporal classification of optical satellite images. Techniques integrating MRFs or CRFs with the aforementioned approaches have also been developed, including combination with neural networks [98], composite kernels [151], theory of evidence [18], and decision fusion [106].

Additional information on multisource data classification can also be found in the recent reviews [61, 119] and in the special issue [143].

7.2 Multilevel Feature Extraction Through Mathematical Morphology

7.2.1 Introduction to Mathematical Morphology

Mathematical Morphology emerged in 1960 s from the pioneering work of Georges Matheron [97] and Jean P.F. Serra [130], who introduced the first formalisms to address the challenges in analysing geometrical structures via transformations and random set modelling with several applications, in particular in the mining industry. Ever since, Mathematical Morphology keeps evolving and has become a well-established discipline in image processing, while the span of application domains that exploit Mathematical Morphology has grown rapidly in recent years. Examples include among others biological and medical image analysis, document processing and remote sensing image analysis. Mathematical Morphology has recently gained an increasing popularity in the remote sensing field that coincided with the increased availability of remotely sensed images with high spatial resolution. Mathematical Morphology provides in this context a series of powerful region-based filtering tools, denoted as connected operators, which are edge-preserving filters that operate by merging connected components or flat zones [126] (i.e., regions composed by iso-level pixels) – see also Chap. 3.

In Mathematical Morphology, connected operators can be implemented by exploiting operators by reconstruction, such as closing and opening by reconstruction. They are filters that perform image transformations by removing or preserving flat zones according to the interaction between an input image and a structuring element, which is a set of neighbouring pixels and is defined according to its shape (e.g., line, circle, etc.) and centre. These operators are based on geodesic transformations and permit the preservation of geometrical characteristics of those objects that are not removed

by the filtering. An alternative approach to efficiently implement connected operators is based on the exploitation of hierarchical representations [127], such as tree-based structures. In a tree representation, regions that compose the image are identified by nodes, which represent the leaves of the tree. In the literature, two main categories of tree representations exist [104]: (i) hierarchies of segmentation (hierarchy of image partitions) and (ii) threshold decompositions (hierarchy of regions). In hierarchies of segmentation, a horizontal cut of the tree results in a set of non-overlapping regions, whose union covers the entire image domain. Examples of representations that belong to this category are minimum spanning tree (MST) [71], alpha-tree [108] (discussed in detail in Chap. 3) and binary partition tree (BPT) [124]. In particular, a BPT allows the image to be decomposed into a collection of regions endowed with suitable inclusion relations, and is constructed and pruned by defining appropriate similarity measures and misclassification rate models [147]. Examples of recent papers using BPTs for hyperspectral and polarimetric SAR image classification include [2] and [147].

In the second category, threshold decompositions, a horizontal cut leads to a set of regions that represent a partial partition. Representations that belong to this category are tree components (i.e., min-tree, max-tree – see also Chap. 3) [68, 69, 125] and tree of shapes (ToS) [27]. A particular family of connected operators that can efficiently be implemented by exploiting threshold decompositions are attribute filters [20, 130]. Attribute filters, being connected operators, have edge-preservation capabilities and perform image simplification by removing flat zones according to a given criterion and attribute. An attribute is any arbitrary measure that can be computed on a flat zone with the goal of describing its geometrical or semantic properties. Given a tree structure, it is possible to compute an attribute for each node of the tree structure. The filtering of the tree is performed by pruning those nodes whose attribute value does not satisfy a predefined criterion, usually defined by a threshold. It is worth mentioning that while the tree structure of an image is fixed, the emerging image simplification may vary depending on the specific attribute chosen for the filtering. The high flexibility of attribute filters and the possibility to perform multi-attribute analysis on a scene allow the extraction of complementary information regarding the spatial arrangement, which can be exploited to improve the discrimination between structures.

The following sections present an overview on recent developments in Mathematical Morphology devoted to the extraction of multiscale features for remote sensing image classification, focusing on connected operators built on tree structures based on threshold decompositions. Such operators are attribute filters computed on tree components and tree of shapes. Note that the background material recalled in the next section about tree-based image representations is also partially covered in Chap. 3, yet it is included here to ensure that the chapter is self-contained.

7.2.2 Theoretical Background

7.2.2.1 Connected Operators

Let a one-channel grey-scale digital image f (see Fig. 7.1a) be defined as a mapping of a set of vertices $V \subseteq \mathbb{Z}^2$ into a set of scalar values $H \subseteq \mathbb{Z}$:

$$f : V \rightarrow H, \tag{7.1}$$

a partition Π_f of f can be defined as a division of the space V into a set π_f of non-empty and disjoint *connected components* (or *flat zones*) CC_i, formally expressed as in [131]:

$$\pi_f = \{CC_1, \ldots, CC_n\}, \tag{7.2}$$

and fulfilling the following property:

$$CC_1, CC_2 \in \pi_f \Longrightarrow CC_1, CC_2 \neq \emptyset \wedge CC_1 \cap CC_2 = \emptyset \quad | \quad \bigcup_i CC_i = V. \tag{7.3}$$

An example of partition is shown in Fig. 7.1c. Let a *flat zone* be a region of connected pixels characterized by the same intensity according to the classic 4 or 8-connectivity rule [133] (see Fig. 7.1b). The connectivity property relies on the concept of *path* in graph theory. The image f can be seen as an undirected graph $G = (V, E)$ where V is a set of vertices representing the pixels, and E is a family of non-ordered pairs of vertices (v_i, v_j) which model the connectivity [133]. A graph G is said to be connected if, for any $p, q \in V$, there exists a path from p to q, which is a sequence of $n > 1$ vertices (i.e., $p = p_1, ..., p_n = q$) such that every $p_i \in V$, and any two successive pixels of the sequence are adjacent $e_{p_i, p_{i+1}} \in E$.

Starting from this definition, a level set of f is defined as $F_h = \{p \in V | f(p) = h\}$ with $h \in H$. At each level set F_h, there may be N connected components $CC_h^k(f)$, with $k \in \{1, ..., N\}$. They are defined as level h components of F (i.e., *flat zones* [126]), and their union within the image f forms a partition Π_f. An operator ψ that acts on an image f is considered connected if it provides a coarser partition π_ψ (i.e., containing less *flat zones*) than the initial one π_f: $\pi_f \sqsubseteq \pi_{\psi(f)}$, meaning that for each pixel $p \in V$, $\pi_f(p) \subseteq \pi_{\psi(f)}(p)$ [105]. Consequently, the regions composing the output partition π_ψ are created by merging the regions of π_f.

Connected components can either be removed or fully preserved, and *connected operators* are therefore edge-preserving operators since they preserve the geometrical detail of the regions that are not processed.

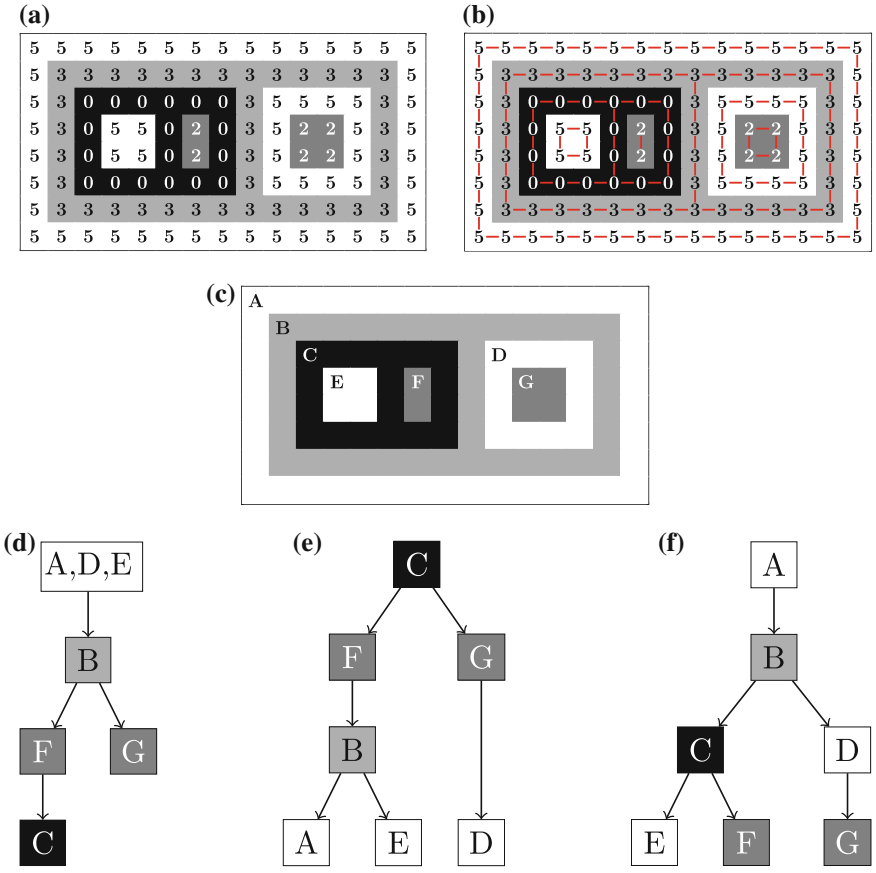


Fig. 7.1 Example of threshold decompositions: **a** Grayscale image with levels of intensity ranging from 0 to 5; **b** 4-connectivity relation induced by the equality of gray levels; **c** partition of **a** into connected components; **d** min-tree; **e** max-tree and **f** tree of shapes (color figure online)

These operators are usually considered as filtering tools; the coarseness of the partition generated (i.e., filtered image) is determined by a size-related filter parameter. One of the most successful implementations of connected operators relies on the tree-based image representations, which are defined next.

7.2.2.2 Threshold Decompositions

A set of partitions π_f of the space V can be organized hierarchically in a tree structure if inclusion relations among the components can be established. For instance, *connected components* can be organized hierarchically: $CC_1, CC_2 \in \pi_f$ are either nested (i.e., $CC_1 \subseteq CC_2$ or $CC_2 \subseteq CC_1$) or not.

Starting from this definition, there are three main threshold decompositions [12] developed in the Mathematical Morphology framework: the component trees (*min-tree* and *max-tree* [68, 125]) and the *tree of shapes* [27]. In the case of component trees, the hierarchy between nodes is driven by an ordering criterion of their grey-levels, whereas, in the case of *tree of shapes*, the ordering criterion follows the inclusion relationship of the regions according to a saturation operator [101], where bright and dark components are simultaneously represented. The *tree of shapes* merges the information of the *min-tree* and *max-tree* into a single structure, leading to a self-dual representation of the image.

7.2.2.3 Component Trees

Component trees were introduced by Jones [68, 69] as efficient image representations that enable the computation of advanced morphological operators in a simple way. They are hierarchical structures that encode the threshold sets and their inclusion relationship, e.g., *min-tree* and *max-tree*, which is shown in Fig. 7.1d, e respectively.

> The *min-tree* models the inclusion of regions according to the grey-level ordering criterion ($\leq$), thus the tree contains only the shapes that are darker than their neighbourhood (i.e., the grey-level of each region is lower than the one of their neighbourhood). The root of the *min-tree* is the entire image domain at the greatest grey-level value, while the leaves are the regional minima. The *max-tree* is dual, and it contains only the regions that are brighter with respect to their neighbouring pixels. In this case, the root is the whole image at the lowest grey-level and the leaves are the regional maxima.

More formally, let the set of scalar values $H \subseteq \mathbb{Z}$ be characterized by an ordering relation $\leq$. For any $h \in H$, a lower $[f \leq h]$ and an upper $[f \geq h]$ threshold sets are defined by:

$$[f \leq h] = \{p \in V | f(p) \leq h\}, \tag{7.4}$$

$$[f \geq h] = \{p \in V | f(p) \geq h\}. \tag{7.5}$$

Let $\mathbb{P}(V)$ be the power set of V, i.e., the set of all the possible subsets of V. Given $X \in V$, the set of connected components of X is denoted as $CC(X) \in \mathbb{P}(V)$. The lower $L_\lambda(f)$ and the upper $U_\lambda(f)$ peak components at level h are determined by:

$$L_h(F) = \{X, X \in CC([f \leq h])\}, \tag{7.6}$$

$$U_h(F) = \{X, X \in CC([f \geq h])\}. \tag{7.7}$$

Finally, the sets of lower $L(f)$ and upper $U(f)$ *connected components* are defined by:

$$L(f) = \cup_h L_h(f), \tag{7.8}$$

$$U(f) = \cup_h U_h(f). \tag{7.9}$$

If $\leq$ is a total relation, any two *connected components* $CC_1, CC_2 \in L(f)$ or $CC_1, CC_2 \in U(f)$ are either nested or not. The inclusion relations of the *connected components* within the set $L(f)$ and $U(f)$ is modelled by the *min-tree* and the *max-tree*, respectively. If $L_h(f) = \{X, X \in CC([f = h])\}$ and $U_h(f) = \{X, X \in CC([f = h])\}$ is the set of N *connected components* at a fixed grey-level $h \in H$, a node of the *min-tree* and *max-tree* represents a unique *connected component* $N_h^k(f)$, with $k \in \{1, ..., N\}$.

Component trees have been widely used for computing attribute filters [20, 145], pattern spectra [109, 145], and multi-scale decompositions [107]. A complete comparison of the different (sequential and parallel) algorithms proposed in the literature for their computation are detailed in [26].

7.2.2.4 Tree of Shapes

The *tree of shapes* (also known as topographic map) is a hierarchical representation of the *connected components* within a grey-level image (i.e., zones enclosed by an isolevel line).

> The *tree of shapes* is self-dual representation since it makes no assumption about the contrast of objects (either light object over dark background or the contrary). The *tree of shapes* can be interpreted as the result of merging the *min-tree* and *max-tree* [125] into a single tree, as shown in Fig. 7.1f.

It was firstly introduced by Monasse et al. [101], where the structure was computed by exploiting the *fast level line transform* (FLLT) algorithm: it first computes the pair of dual component trees and then obtains the *tree of shapes* by merging them. Caselles et al. [28] introduced the *fast level set transform* algorithm (FLST), which is based on a region-growing approach to decompose the image. An operation called *saturation* is applied to the *connected components* aiming at filling holes, resulting in flat regions (or *shapes*) obtained by progressively merging nested regions. Specifically, the algorithm extracts each branch of the tree starting from the leaves and growing them up to the root until only a single flat region is obtained. Song et al. [135], proposed to retrieve the *tree of shapes* by building the tree of level lines and exploiting the interior of each level line. Recently, Geraud et al. [59] proposed a new algorithm to compute the *tree of shapes* in order to reduce the computational complexity and overcome the restriction to only 2-D images of the previous methods. The algorithm computes the *tree of shapes* with quasi-linear time complexity when data quantification is low (typically 12 bits or less) and it works for n-D images. The first parallel algorithm to compute the *tree of shapes*, which is based on the algorithm proposed in [59], is presented in [35].

More formally, given the set $X \in V$, let ∂X be the border of X and $\bar{X}$ the complementary of X. The hole-filling operator $H : \mathbb{P}(V) \rightarrow \mathbb{P}(V)$ is defined by:

$$H(X) = V \setminus CC(\bar{X}, \partial X), \tag{7.10}$$

where $CC(\bar{X}, \partial X)$ is the *connected component* of $\bar{X}$ linking with the image border. Given the operator H, a shape is any element of the set:

$$S = \{H(L)\}_h \cup \{H(U)\}_h. \tag{7.11}$$

If $\leq$ is total, any two shapes are either disjointed or nested, hence the cover of $S, \subseteq$ makes the *tree of shapes*.

The definition of the shapes as hole-filled *connected components* of the lower $L(f)$ and upper $U(f)$ threshold set proves that the *tree of shapes* can be seen as a merge of the *min-tree* and *max-tree*. However, the hole-filling operation creates shapes within neither the *min-tree* nor the *max-tree*.

7.2.2.5 Attribute Filters

Component trees and *trees of shapes* are very attractive since they allow edge-preserving operations. Accordingly, among different types of classical morphological operators, originally developed in [130, 131], *attribute filters* [20] have been largely diffused.

Attribute filters are connected operators that act on *connected components* according to an *attribute* criterion. By *attribute* we mean any measure that could be computed on a *connected component*.

Therefore, an *attribute* can be related to the geometry and shape (e.g., area, bounding box, image moments), to the texture (e.g., standard deviation, entropy), contour and context (such as the context-based energy estimator [154]). In detail, given a tree representation, the value of an *attribute* A is evaluated on each node $N_h^k(f)$ and compared with a reference threshold λ in a binary predicate T_λ:

$$T_\lambda := A\left(N_h^k(f)\right) \geq \lambda. \tag{7.12}$$

In general terms, if the predicate is true, the node is preserved, otherwise it is removed. In literature, different strategies of remove/preserve decisions have been defined [145]. According to the type of predicate and the property of A, the resulting connected operator can be defined as increasing or non-increasing. In the context of a tree structure, this characteristic is related to the criterion assessed for

each node. An operator is considered increasing when the predicate is in the form $T_\lambda = A(N_h^k(f)) \geq \lambda$ or $T_\lambda = A(N_h^k(f)) \leq \lambda$ and the attribute is increasing, meaning that the attribute computed on a node is always bigger than those computed on its descendant nodes in the tree. An example of increasing attributes are those that in some way are related to the scale of the region, such as the *area*, defined as the number of pixels composing the connected region. Vice versa, when the attribute is not increasing, the attribute value computed on a node can be smaller than those computed on its descendants in the tree. Examples of non-increasing attributes are those related to the analysis of shape, texture, energy, such as standard deviation and moment, which are independent of the dimension of the connected component.

Regarding the filtering strategies, two general approaches might be used: pruning and non-pruning strategies. The former strategy consists in removing entire branches of the tree. A single cut is made along each path from leaf to root, and all nodes leaf-side of the cut are collapsed onto the highest surviving ancestor. This strategy is simple to apply especially when the chosen attribute is increasing since all nodes for which the criterion is not verified are organized in entire branches (i.e., if a node has to be removed, all of its descendants also have to be removed). Examples of pruning strategies are the *min* and *max* rules [146]. Non-pruning strategies provide solutions for such cases where the simplification approach is not straightforward, as it is for non-increasing attributes, where the descendants of a node to be removed have not necessarily been removed). For instance, the simplification of the tree is not limited to the removal of entire branches but also isolated nodes might be removed along a root path. Many approaches have been proposed in the literature, such as the Viterbi algorithm [54, 125], optimization methods [155], the *direct* rule [146] and the *subtractive* rule [145].

7.2.3 Multilevel Image Representation

The joint use of spectral and spatial information has become a standard procedure in image classification, especially in high resolution remote sensing image analysis, where the extraction of contextual information allows us to deal with the increase of within-class spectral variability introduced by the high geometrical detail (see also Chap. 3).

> An efficient strategy for the extraction of contextual information is represented by multilevel analysis, where the spatial relations between pixels belonging to homogeneous regions or structures are modelled at different levels of abstraction, from pixel-level (i.e., high resolution) to region-level (i.e., low-resolution).

Considering the exploitation of both increasing and non-increasing attributes, the term "multiscale" needs to be replaced by the more general term "multilevel." Starting

from the definition of *attribute filters*, a multilevel representation of an image can be obtained by performing a sequential morphological filtering considering coarser filtering thresholds, resulting in a vector of filtered images, denoted as *profile*. More formally, given a family of L either increasing or decreasing criteria $T = \{T_\Lambda\}$, where $\Lambda = \{\lambda_i\}_{i=1}^{L}$ identifies a set of scalar values used as reference thresholds, a *profile* is defined as a vector of filtered images resulting from a sequential application of a connected operator ψ, where a criterion T_{λ_i} is evaluated at each filter step:

$$P_\psi := \left\{\psi^{\lambda_i}\right\}_{i=1}^{L}. \tag{7.13}$$

Such structure was introduced in [116], where a *morphological profile* was defined as the concatenation of the *morphological closing profile* and *morphological opening profile* obtained by exploiting operators by reconstruction. In this case, the use of operators by reconstruction, which are based on a SE and thus increasing operators, led to a multiscale decomposition of the scene. In [40], the same concept was applied to *attribute filters*, introducing the *attribute profiles*. By considering the *max-tree* and the *min-tree* representations, the computed *attribute opening profile*, P_γ, and *attribute closing profile*, P_ϕ, are respectively defined as:

$$P_\gamma = \left\{\gamma^{T_0}, \gamma^{T_{\lambda_1}}, \ldots, \gamma^{T_{\lambda_L}}\right\}, \tag{7.14}$$

$$P_\phi = \left\{\phi^{T_0}, \phi^{T_{\lambda_1}}, \ldots, \phi^{T_{\lambda_L}}\right\}, \tag{7.15}$$

where γ^T and ϕ^T represent the *attribute opening* and *attribute closing*, respectively, $\{T_i\}$ is a criterion evaluated on the set of thresholds Λ and $\phi^{T_0}(f) = \gamma^{T_0}(f) = f$ represents the original image. An *attribute profile (AP)* is defined as a vector of filtered images resulting from the concatenation of the *attribute closing profile* taken in reverse order, P_ϕ^-, such that each entry is greater than or equal to the subsequent one, with the *attribute opening profile*:

$$AP = \left\{P_\phi^- / \phi^{T_0}, P_\gamma\right\}. \tag{7.16}$$

The resulting multilevel structure is composed of $2L + 1$ filtered images, given by L closings, the original image and L openings). Analogously, when considering the contrast invariant operator ρ based on the *tree of shapes*, the profile P_ρ, named *self-dual attribute profile* [30, 41], is obtained as:

$$P_\rho = \left\{\rho^{T_0}, \rho^{T_{\lambda_1}}, \ldots, \rho^{T_{\lambda_L}}\right\}, \tag{7.17}$$

with $\rho^{T_0}(f) = f$. In the case of *self-dual attribute profile (SDAP)*, the resulting feature vector counts $L + 1$ filtered images.

7.2.4 Multi-channel and Multi-attribute Representations

As *attribute filters* and their multilevel representations are defined for 2D images, any of their extensions to a multivariate scenario is an ill-posed problem. A simple strategy to extract multilevel representations in a multi-channel scenario relies on performing image decompositions considering each channel separately, obtaining a vector of profiles. However, such strategy is feasible when the number of channels is relatively small, while it becomes unattainable when high dimensional data are considered, as it is in case of hyperspectral data (see Chap. 2). To mitigate this issue and to be able to extract features to model the contextual information in high-dimensional data, in [9], *morphological profiles* were built on a sub-set of features obtained via principal component analysis (PCA). The concatenation of each *morphological profile* provided a new structure named *extended morphological profile (EMP)*. A similar strategy was adopted to create multi-channel structures exploiting connected operators based on tree representations. In [39], *extended attribute profiles (EAPs)* were introduced as the concatenation of *attribute profiles*:

$$EAP = \Big\{AP(C_1), AP(C_2), \ldots, AP(C_R)\Big\}. \tag{7.18}$$

As a general strategy, a subspace of R features, C_i, with $i = 1, \ldots, R$, is extracted from the original high-dimensional data via feature dimensionality reduction and used as input to the morphological analysis. In literature, several studies can be found focusing on the impact of different feature extraction approaches. In [41, 50], independent component analysis (ICA) was considered as feature extraction for dimensionality reduction. Comparison of *extended attribute profiles* built on subsets of features extracted by using PCA, kernel-PCA, discriminant analysis feature extraction (DAFE), decision boundary feature extraction (DBFE) and non-parametric weighted feature extraction (NWFE) [76] can be found in [11, 36, 96, 114]. Wavelet transform was also exploited in the context of hyperspectral image classification in [120].

Due to the higher flexibility in modelling contextual information, such operators have been proven to be more effective in image classification with respect to morphological filters based on structuring elements and their extensions. In case of self-dual operators, the multi-channel extension of *self-dual attribute profiles* were introduced in [30], resulting in the definition of *extended self-dual attribute profiles (ESDAPs)*:

$$ESDAP = \Big\{SDAP(C_1), SDAP(C_2), \ldots, SDAP(C_R)\Big\}. \tag{7.19}$$

Considering the flexibility of *attribute filters* in modelling the filter criterion, a further extension is defined by the concatenation of *extended profiles* obtained by considering different attributes, resulting in the definition of *extended multi-attribute profile (EMAP)* [39], which is defined as follows:

$$EMAP = \left\{ EAP_{A_1}, EAP_{A_2}, \ldots, EAP_{A_Q} \right\}, \tag{7.20}$$

where A_i represents the i-th chosen attribute, with $i = 1, 2, \ldots, Q$. In order to avoid multiple presence in the final structure, the original input feature C_1 is included once only in the first EAP.

Both *EAPs* and *ESDAPs* (and their multi-attribute extensions) can be seen as rich descriptors of the contextual domain of the investigated scene. On the other hand, filters based on dual representations process regions independently on the contrast information. This characteristic makes the *ESDAPs* more versatile in modelling the spatial context [37] as compared to non-dual operators, especially for those images characterized by a prevalence of dark or bright regions.

7.2.5 Automatic Filter Parameter Selection

Suitable sets of filter parameters (i.e., thresholds) need to be identified to extract multilevel features able to represent the contextual domain of the scene under investigation. Empirical searching based on field-knowledge and visual inspection represents the common approach, often requiring multiple filtering tests, causing unfavourable effects in terms of time and computational efficiency. Automatic approaches aimed at minimizing the manual intervention were recently proposed in literature. In [90], a large vector of attribute values was derived by computing a chosen attribute on objects extracted from a preliminary classification of the investigated scene. The final threshold set was obtained by clustering the vector of values and selecting the minimum attribute values of each cluster as a candidate threshold. The method provided better or similar results to the manual selection. In [95], the threshold set was automatically identified based on a statistical analysis of the available training samples. The procedure was defined for *attribute profiles* based on the *standard deviation* attribute in a supervised classification scenario. Due to the high dimensionality of the obtained profiles, a further dimensionality reduction procedure was required in order to avoid the raising of the Hughes' phenomenon (see Chap. 2). In [56], the automatic selection was based on the analysis of the characteristic function of the pattern spectrum [92, 145], which corresponds to the probability density function of the granulometric curve of the attribute profile, i.e., a curve related to the size distribution of the structures in the image [133]. The selected thresholds were those whose characteristic function best approximated the one obtained by considering a larger set of thresholds. The method required an initial set of thresholds, which was manually defined prior to the filtering.

Aiming at exploiting the inner properties of hierarchical representations of the image, the authors in [29] have proposed a methodology for the automatic selection of thresholds based on the concept of granulometric characteristic functions (GCFs) [29], suitable for representations obtained by using *components trees* and *trees of shapes*.

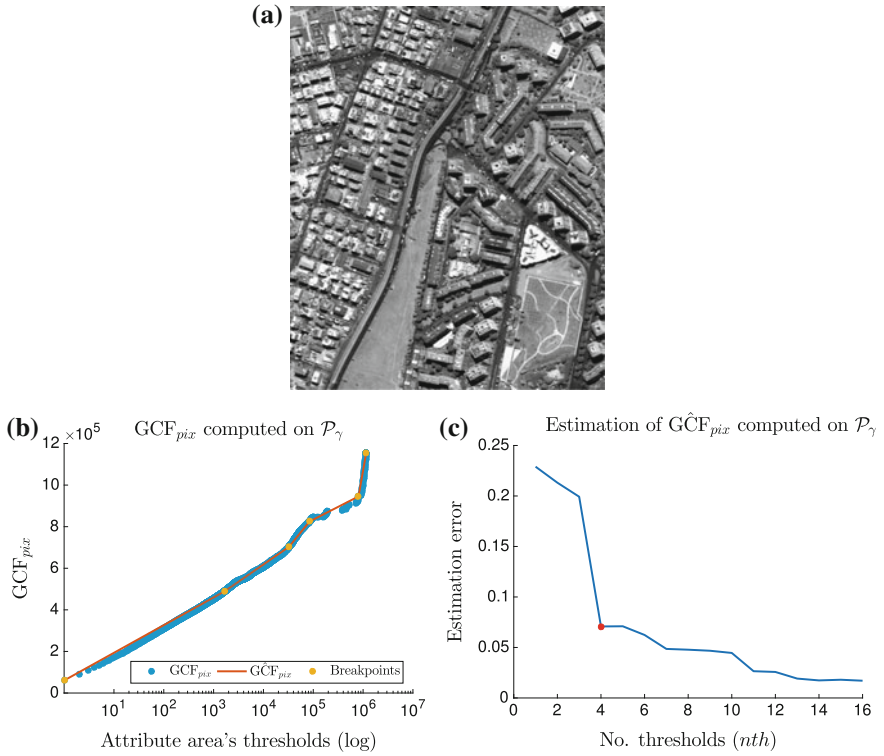


Fig. 7.2 Example of a GCF and threshold selection. **a** Rome data set (panchromatic channel). **b** GCF derived for Rome data set considering the number of changed pixels as measure of interest (GCF$_{pix}$). The GCF$_{pix}$ is computed on the max-tree representation considering the attribute *area*. The estimated GĈF is represented by the red line, while the *yellow circles* identify the breakpoints, which are used to derive the final set of thresholds. **c** Estimation error showing the number of threshold chosen (*red point*)

GCFs are descriptive functions that derive directly from the tree representation of the image and are used to describe a particular behaviour of the multi-level decomposition according to a global measure of interest.

Following this definition, a GCF can be formulated as:

$$\mathrm{GCF}(P_\psi(f)) = \{M(\psi_i)\}_{i=1}^{L}, \tag{7.21}$$

where $M(\psi)$ represents a measure of interest computed at level (threshold) i. In particular, three measures were defined:

- *Sum of grey-level values*: Similarly to the conventional granulometry, this measure quantifies the global change in grey-level caused by the filtering.

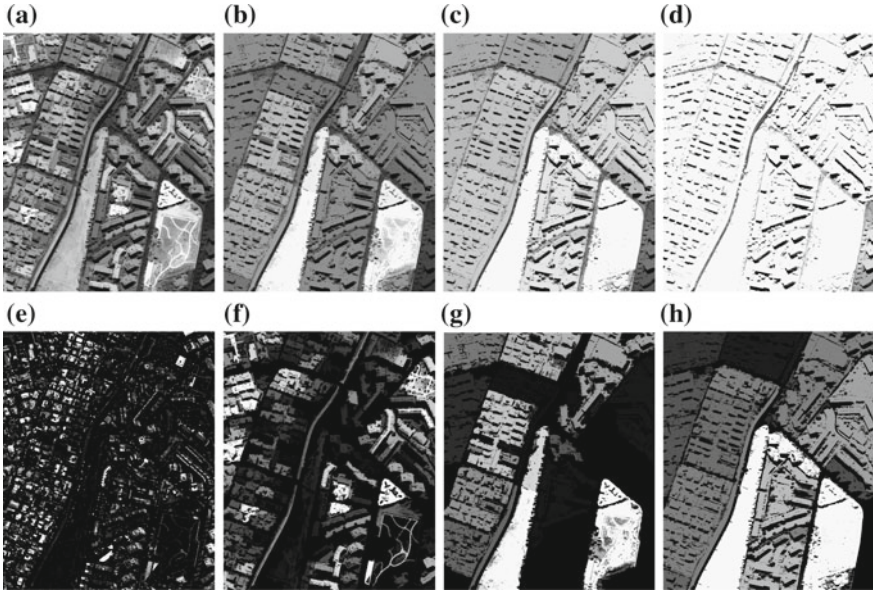


Fig. 7.3 Example of opening attribute profile (based on max-tree representation) computed for the Rome data set by using the attribute *area*. The thresholds are obtained by analysing the GCF_{pix} depicted in Fig. 7.2b obtained by considering the number of changed pixels as measure of interest: **a** $\lambda = 1658$; **b** $\lambda = 32691$; **c** $\lambda = 84485$; **d** $\lambda = 809471$. **e**–**h** Residual between adjacent features (DAP)

- *Number of changed pixels*: This measure quantifies the filtering effect in terms of total number of pixels that have changed grey-value. The obtained GCF is more sensitive to changes in the spatial extent of the regions rather than in grey-levels.
- *Number of changed regions*: This measure quantifies the filtering effect in terms of total number of connected components that are affected. This measure is topologically invariant to both the spatial extent and grey-level variations induced by the filtering.

An example of a GCF computed according to the number of changed pixels and derived from a max-tree representation is shown in Fig. 7.2b. For this example, the Rome data set represented by a panchromatic image (see Fig. 7.2a) is considered. The *max-tree* representation of the image counts 268004 nodes with 5337 unique values for the attribute *area*, representing the whole set of possible thresholds. The GCF is obtained by computing the measure of interest for the whole set of values.

Once the GCF is computed, an iterative procedure is performed to identify the subset of thresholds that best approximate the global behaviour of the GCF. In particular, at each iteration the approximated, $\hat{GCF}$ (see Fig. 7.2b) is computed by employing a piecewise linear regression model [80] for an increasing number of thresholds and evaluated in terms of estimation error. To this purpose, the normalized root mean squared error (NRMSE) is used, shown in Fig. 7.2c. Considering the monotonically

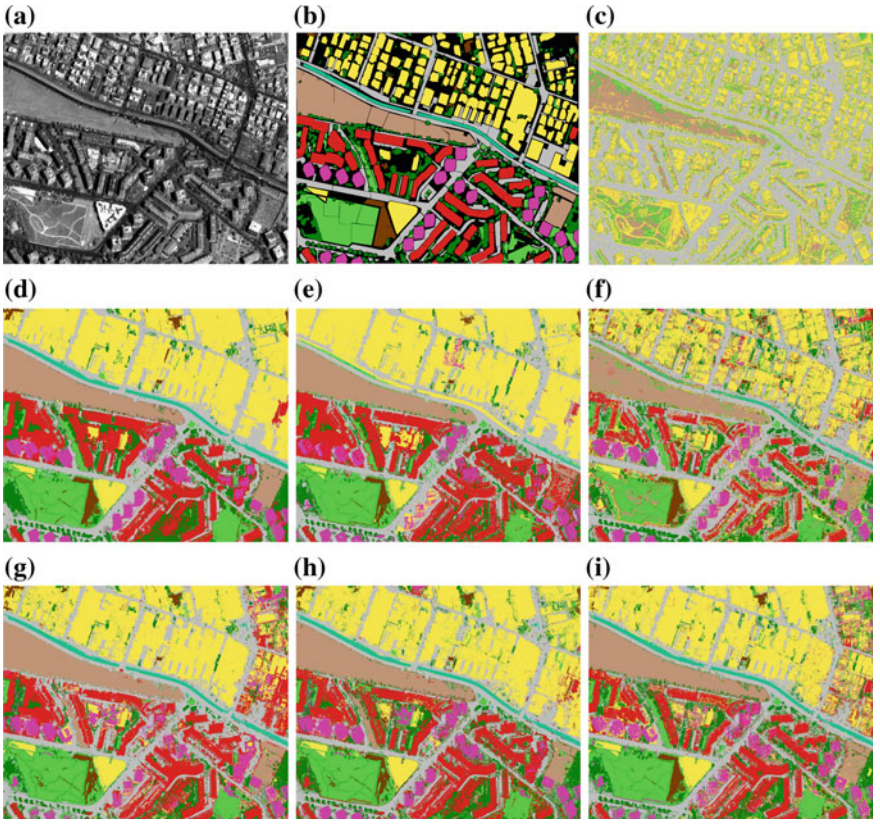


Fig. 7.4 Classification results for Rome data set: **a** original panchromatic image; **b** reference data: buildings, apartment blocks, roads, railway, short vegetation, trees, bare soil, soil, towers; **c** classification map obtained by using the panchromatic channel alone; **d–f** results obtained by using the attribute *area*; **g–i** results obtained by using the attribute *standard deviation*

decreasing behaviour of the estimation error, the algorithm stops when the point of maximum curvature of the NRMSE function is found, which provide information on the number of thresholds that will compose the final threshold set. The subset of threshold values is then derived by the segments that constitute the $\hat{GCF}$ (see Fig. 7.2b), excluding the extreme values, which correspond to the original input image and to the final filtered image with a constant grey-scale value, respectively. An example of *opening attribute profile* and relative *differential attribute profile* (DAP), which shows the residual between adjacent features due to the filtering, are depicted in Fig. 7.3, showing an effective multi-level representation of the original input scene.

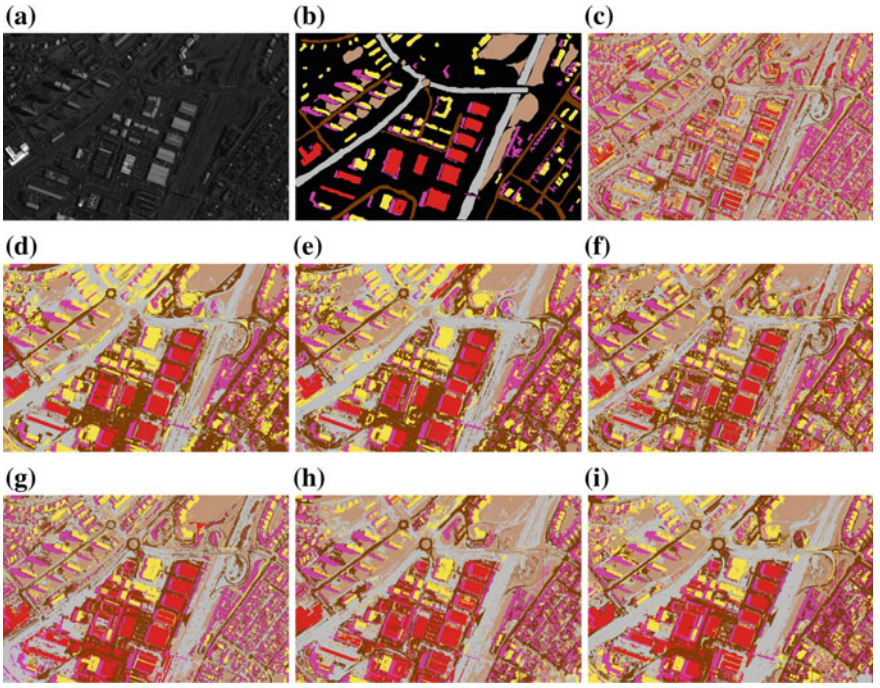


Fig. 7.5 Classification results for Reykjavik data set: **a** original panchromatic image; **b** reference data: small buildings, open areas, shadows, large buildings, large roads, streets; **c** classification map obtained by using the panchromatic channel alone; **d–f** results obtained by using the attribute *area*; **g–i** results obtained by using the attribute *standard deviation*

7.2.6 Experimental Study

In this section, we present experimental results of image classification based on multilevel feature extraction on two high resolution data sets:

D1 The data set is composed of a panchromatic image acquired by the QuickBird satellite sensor on 19th July 2004 over the city of Rome, Italy. The data size is 1188 × 973 pixels with spatial resolution of 0.65 m. The acquired scene is a dense heterogeneous urban area, which includes nine ground reference classes, namely: *buildings*, *apartment blocks*, *roads*, *railway*, *short vegetation*, *trees*, *bare soil*, *soil*, *towers*. The data set and the related reference map are shown in Fig. 7.4a, b respectively.

D2 The data set is composed of a panchromatic image acquired by the IKONOS satellite sensor on 9th August 2001 over the city of Reykjavik, Iceland. The data size is 975 × 639 pixels with spatial resolution of 1 m. For this scene, six classes of interest are defined: *small buildings*, *open areas*, *shadows*, *large buildings*,

Table 7.1 Classification results obtained for the Rome data set. Each profile is built on the panchromatic image considering the attribute

	Panchromatic	P_ρ *area*		
		GCF_{val}	GCF_{pix}	GCF_{reg}
No features	1	6	3	4
Buildings	57.97% (1.31)	95.34% (0.11)	92.95% (0.23)	81.80% (0.24)
Apartment blocks	1.47% (0.92)	87.45% (0.36)	85.34% (0.78)	66.86% (0.66)
Roads	86.87% (0.85)	84.10% (0.53)	80.56% (0.31)	86.48% (0.34)
Railway	0% (0)	91.75% (0.22)	30.40% (2.97)	57.78% (0.68)
Short vegetation	22.31% (4.41)	82.38% (0.25)	77.70% (0.31)	68.64% (0.52)
Trees	1.06% (0.88)	50.71% (0.52)	35.21% (0.76)	49.37% (0.31)
Bare soil	69.64% (2.14)	97.81% (0.10)	89.05% (0.90)	88.21% (0.77)
Soil	0% (0)	73.66% (0.89)	56.48% (1.43)	74.45% (1.00)
Towers	0.44% (0.19)	79.96% (0.76)	58.53% (1.96)	76.71% (0.34)
AA	26.64% (0.18)	82.57% (0.14)	67.36% (0.32)	72.26% (0.12)
OA	41.53% (0.06)	84.27% (0.06)	76.59% (0.15)	75.41% (0.04)
k	28.06% (0.12)	81.25% (0.07)	71.90% (0.19)	70.63% (0.05)
		P_ρ *standard deviation*		
		GCF_{val}	GCF_{pix}	GCF_{reg}
No features		4	5	7
Buildings		86.87% (0.77)	92.54% (0.16)	88.86% (0.24)
Apartment blocks		77.52% (1.24)	80.89% (0.53)	79.41% (0.48)
Roads		82.82% (0.72)	80.33% (0.27)	82.36% (0.29)
Railway		90.51% (0.56)	91.27% (0.43)	91.56% (0.21)
Short vegetation		73.67% (0.30)	76.46% (0.33)	85.22% (0.21)
Trees		34.56% (1.30)	47.36% (0.69)	54.62% (0.46)
Bare soil		95.26% (0.41)	96.05% (0.27)	96.04% (0.15)
Soil		58.90% (1.36)	63.35% (0.96)	76.30% (0.58)
Towers		60.82% (1.58)	74.28% (0.50)	66.83% (0.33)
AA		73.44% (0.25)	78.06% (0.14)	80.13% (0.12)
OA		76.18% (0.17)	80.22% (0.05)	81.02% (0.09)
k		71.53% (0.21)	76.44% (0.06)	77.39% (0.11)

large roads and *streets*. The data set and reference map are shown in Fig. 7.5a, b respectively.

The experimental analysis focuses on the use of *self-dual attribute profiles* for image classification. *Self-dual attribute profiles* are derived by *tree of shape*, which are self-dual connected operators that act simultaneously on bright and dark regions. In literature, several studies can be found focusing on the impact of different feature extraction approaches. In [41, 50], ICA was considered as feature extraction for dimensionality reduction. Comparison of *EAPs* built on subsets of features extracted

Table 7.2 Classification results obtained for the Reykjavik data set. Each profile is built on the panchromatic image considering the attribute

	Panchromatic	P_ρ *area*		
		GCF_{val}	GCF_{pix}	GCF_{reg}
No Features	1	5	6	4
Small buildings	39.18% (1.72)	81.45% (0.53)	81.94% (0.46)	74.67% (0.64)
Open areas	47.27% (1.34)	77.06% (1.12)	77.18% (0.85)	66.72% (0.91)
Shadows	93.62% (0.69)	94.61% (0.61)	94.74% (0.38)	94.25% (0.46)
Large buildings	45.42% (1.82)	92.48% (0.40)	93.38% (0.37)	76.73% (1.32)
Large roads	61.25% (1.33)	88.29% (3.11)	90.04% (1.54)	72.37% (2.49)
Streets	39.14% (0.54)	72.55% (1.11)	77.81% (1.15)	66.25% (2.44)
AA	54.31% (0.14)	84.41% (0.60)	85.85% (0.19)	75.17% (0.75)
OA	52.62% (0.13)	83.57% (0.63)	85.04% (0.20)	73.88% (0.76)
k	42.75% (0.16)	80.17% (0.76)	81.95% (0.25)	68.45% (0.92)
		P_ρ *standard deviation*		
		GCF_{val}	GCF_{pix}	GCF_{reg}
No Features		4	4	6
Small buildings		58.54% (1.01)	62.64% (1.10)	73.98% (0.46)
Open areas		55.74% (2.33)	71.09% (1.56)	79.16% (1.10)
Shadows		92.75% (0.19)	92.54% (0.54)	93.81% (0.29)
Large buildings		84.00% (0.99)	83.88% (0.57)	91.72% (0.40)
Large roads		75.72% (3.71)	81.01% (1.60)	87.46% (0.58)
Streets		49.84% (2.82)	57.05% (0.65)	63.11% (1.93)
AA		69.43% (0.18)	74.70% (0.16)	81.54% (0.18)
OA		67.70% (0.16)	73.75% (0.15)	80.84% (0.16)
k		60.99% (0.20)	68.25% (0.19)	76.84% (0.20)

by using PCA, kernel-PCA, DAFE, DBFE and NWFE [76] can be found in [11, 36, 96, 114]. Wavelet transform was also exploited for hyperspectral image classification in [120]. Further experimental analysis that focus on the exploitation of *attribute profiles* can be found in [8, 38, 40, 152]. Several strategies have also been defined for improving the spatial information extracted from *attribute profiles* and their extensions [46, 50, 134, 159].

Self-dual attribute profiles are extracted by employing the automatic filter parameter selection and considering the three measures defined in Sect. 7.2.5. The obtained multilevel representation is then used as input to a supervised learning algorithm. An ensemble random forest classifier was chosen as supervised learning algorithm, in which the minimum out-of-bag (OOB) error was exploited to estimate the optimum number of trees. For each experiment, a tenfold cross-validation was performed and means and standard deviations of class accuracies, overall accuracies (OA), average

accuracies (AA) and kappa coefficients (k) are reported in Tables 7.1 and 7.2. Training sets were built based on random selection to be 10% of the reference samples, while the remaining samples were used as test sets.

For the Rome data set, it can be seen that some of the classes of interest can not be discriminated when the panchromatic image is used alone (see Table 7.1). In particular, the classes *apartment blocks*, *railway*, *trees*, *soil*, and *towers* are poorly detected. This is mainly due to insufficient spectral information that characterizes panchromatic images. On the other hand, the use of spatial features to enrich the feature space allows a better classification performance. From the quantitative results shown in Table 7.1, the best classification accuracy was obtained by considering the attribute *area* and selecting GCF_{val} as global descriptor for the image decomposition (see Fig. 7.4d. In the case of Reykjavik data set, the highest classification accuracy was obtained by using the attribute *area* and GCF_{pix} as global descriptor (see Fig. 7.5e). It can be seen in Table 7.2 that some of the combinations of attribute-GCF provide lower classification results. This is due to the fact that GCFs are global descriptors of certain properties of the multiscale image decomposition, which is highly scene dependent. Therefore, for a given attribute, the resulting GCF can prove more suitable than other GCFs. This in turn improves the attribute profile, since it contains salient structures of the scene under investigation.

7.3 Hierarchical Markov Models for Multisource Image Classification

In this section we develop a general method for multitemporal, multiresolution, and multisensor classification of remotely sensed images based on a hierarchical Markovian model. The general concepts of MRF models have been discussed in Chap. 4. Here, we build on a Markovian formulation to fuse multitemporal, multiresolution, and multisensor data for classification purposes. In particular, our objective is to satisfy two major requirements: (i) the method should be parallelizable to handle large amounts of input data, and (ii) the method should provide a structure simplifying the interactions between the images in the input data set. Parallel multi-grid (or pyramidal) schemes are one of the possible approaches satisfying the first requirement.

The hierarchical pyramid structure is a type of signal representation in which images are organized according to their resolutions (see Fig. 7.6). Specifically, a *pyramid* $\mathscr{P}$ is a stack of images $\mathscr{I}_n$ with the scale parameter $n \in \{0, 1, \ldots, R\}$ and R denotes the height of the pyramid. An element of this pyramid is called a *node* and may correspond to a pixel or a group of pixels.

To address the second requirement, for each node of the pyramid we define a set of links to other nodes to model interscale interactions. The theory of multiscale signals has been widely studied [91], and their representations naturally lead to tree-

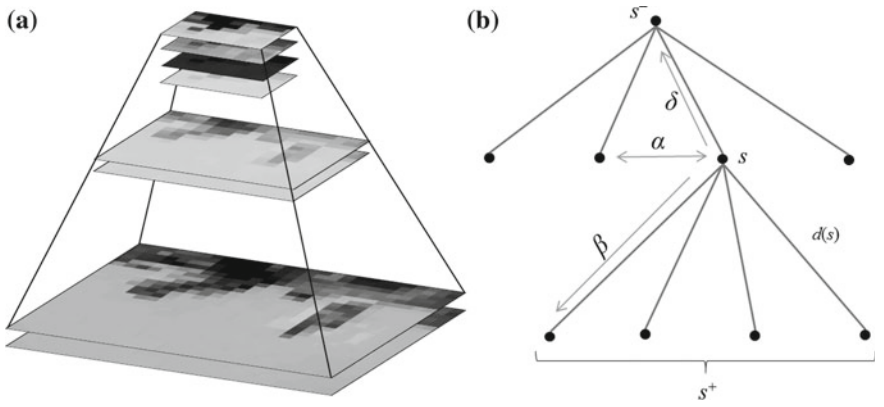


Fig. 7.6 Hierarchical pyramid structure for image processing: **a** images are organized according to their resolutions in a *pyramid* structure; **b** *upward* shift (δ), *forward* shift (β), and interchange (α) operators associated with the relations among a node s in the *pyramid*, its parent s^-, its children s^+, and its descendants $d(s)$

based models [88]. In particular, *dyadic trees* and *quad-trees* have been considered for modeling these inter-scale interactions in 1D and 2D signals, respectively [88]. The selection of these structures is justified by their *causality* properties over scale and by the availability of fast optimization methods. Specifically, we refer to the Markovian causality with respect to scale, i.e., the probability of observing a label at any tree node conditioned on the labels of all the node's ancestors reduces to the dependence on the label of the parent node.

Let us denote a generic node of a quad-tree as s and the set of all nodes as S ($s \in S \subseteq \mathbb{Z}^2$). Each node is a pixel in an image located at the corresponding level of the tree. The set of nodes is then hierarchically partitioned, (i.e., $S = S^0 \cup S^1 \cup \ldots \cup S^R$) where S^n indicates the subset of nodes associated with the nth level ($n = 0, 1, \ldots, R$); $n = R$ denotes the *root* of the tree (coarsest resolution) and $n = 0$ indicates its *leaves* (finest resolution). In this structure, a parent-child relationship can be defined: an *upward shift operator* δ such that $s^- = \delta(s)$ is the parent of node s. The operator δ is not one-to-one, but four-to-one because each parent has four offspring in the quad-tree structure. We define the *forward shift operator* β such that $s^+ = \beta(s)$ is the set of the descendants of s, the *interchange operator* α is defined as between the nodes in the identical scale, and $d(s)$ is the set including s and all its descendants as illustrated in Fig. 7.6. This framework allows data from sensors with different resolutions and spectral band compositions to be integrated in the same processing mainframe.

A novel aspect of the approach described here is the *multitemporal* component (see also Chaps. 8 and 9). We employ multiple pyramids and quad-trees in a cascade, each pyramid being associated with the set of images available at a specific date to characterize the temporal correlations associated with distinct images in the input time series. We build new operators to link between the nodes across different dates.

Therefore, we define a *multitemporal upward shift operator* ω such that $s^{=} = \omega(s)$ is the parent of node s in the previous date of the time series. Furthermore, to characterize the temporal correlation between images given at different dates, we define a *multitemporal interchange operator* τ such that $s^{\#} = \tau(s)$ is the node in the same scale and position as s but in the previous date.

The proposed multitemporal hierarchical structure allows supporting the joint classification of both multitemporal and multiresolution images. In case when only one image is available at one of the acquisition dates, then such image is included in the finest resolution layer of the corresponding quad-tree. If multiple images with distinct resolutions are available, then the images are included at different quad-tree levels. In this case, the quad-tree underlying assumption requires that the resolutions of the input images are related as powers-of-2. This condition is satisfied with minor approximations by most current multiresolution spaceborne optical sensors, so it is operatively only a mild restriction. After the input images are included in the layers corresponding to their resolutions, level 0 of the quad-tree corresponds to the finest-resolution input image, while some intermediate levels may generally lack data and are filled using wavelet transforms [91] of the images on the lower (finer resolution) layers.

The hierarchical structure allows, in a natural way, the use of an explicit statistical model through a hierarchical MRF. This formulation is based on a set of *random fields*, which are associated with the different scales, and exploiting the operators defined above on the quad-tree structure. Let us denote the class label of site s as a discrete random variable x_s. If there are M classes in the considered classification scenario, then the labels take values from the set $\Lambda = \{0, 1, \ldots, M-1\}$, $x_s \in \Lambda$. The class labels of all pixels can be collected in a set $\mathbf{X} = \{x_s\}_{s \in S}$ of random fields $\mathbf{X}_t^n = \{x_s\}_{s \in S_t^n}$ associated with each scale n and time t, where S_t^n is the related set of lattice points. The configuration space $\Omega = \Lambda^{|S|}$ is the set of all global discrete labellings (i.e., $\mathbf{X} \in \Omega$).

A *hierarchical Markov model* on the quad-tree structure is determined by the following assumptions:

(i) The sequence of random fields from coarse to fine scales forms a Markov chain over scale and time:

$$P\left(\mathbf{X}_t^n | \mathbf{X}_p^q,\, p < t, q > n\right) = P\left(\mathbf{X}_t^n | \mathbf{X}_t^{n+1}, \mathbf{X}_{t-1}^{n+1}\right)$$

(ii) The transition probabilities of this Markov chain factorize so that the components (nodes) of $\mathbf{X}_t^n$ are mutually independent given $\mathbf{X}_t^{n+1}$ and $\mathbf{X}_{t-1}^{n+1}$:

$$P\left(\mathbf{X}_t^n | \mathbf{X}_t^{n+1}, \mathbf{X}_{t-1}^{n+1}\right) = \prod_{s \in S_t^n} P\left(x_s | x_{s^-}, x_{s^=}\right).$$

The quad-tree structure benefits from causality and parallelizability properties discussed above, but also allows non-iterative algorithms. The latter results in a decrease of computational time compared to iterative optimization required for graphs [75].

7.3.1 Bayesian Classification

The aim of the classification is to estimate the value of the hidden label field $\mathbf{X}$ given a realization of a random field of observations $\mathbf{Y} = \{\mathbf{y}_s\}_{s \in S}$ attached to the set of nodes S. In this context, we consider the problem of inferring the "best" configuration $\hat{\mathbf{X}} \in \Omega$. The standard Bayesian formulation of this inference problem consists of minimizing the appropriate risk functions:

$$\hat{\mathbf{X}} = \arg\min_{\mathbf{X}' \in \Omega} E\left\{C(\mathbf{X}, \mathbf{X}')|\mathbf{Y}\right\}, \tag{7.22}$$

where C is the cost function penalizing the discrepancy between the estimated configuration and the "ideal" random configuration, and $E\{\cdot\}$ is the expectation operator.

Among the different classification algorithms employed on a quad-tree structure in the literature, two have been widely used. The first algorithm aims to estimate exactly the *maximum a posteriori (MAP)* configuration with the following cost:

$$C_{MAP}(\mathbf{X}, \mathbf{X}') = 1 - \delta(\mathbf{X}, \mathbf{X}') = 1 - \prod_{s \in S} \delta\left(x_s, x'_s\right), \tag{7.23}$$

where $\delta(\cdot)$ is the Kronecker delta (i.e., $\delta(a, b) = 1$ for $a = b$, and $\delta(a, b) = 0$ otherwise). This function implies the same cost for all pairs of configurations that differ in, at least, one site. From Eqs. (7.22) and (7.23), the MAP estimator of the label field is:

$$\hat{\mathbf{X}}_{MAP} = \arg\min_{\mathbf{X} \in \Omega} P(\mathbf{X}|\mathbf{Y}) \tag{7.24}$$

This combinatorial optimization problem can be resolved with: (i) a Kalman-like filter, due to the formal similarity between MRF models and the spatio-temporal models used in Kalman approaches for optical flow [87], or (ii) a Viterbi algorithm [54]. The extension of the Viterbi algorithm, which computes the exact MAP estimate of $\mathbf{X}$ given $\mathbf{Y}$ on the quad-tree has been first introduced in the context of probabilistic expert systems [44], and then in the context of image classification by proposing a non-iterative algorithm on the quad-tree. However, these algorithms are affected by underflow problems because of the small probabilities involved [115]. Moreover, according to (7.23), the MAP cost function penalizes the discrepancies between configurations regardless of their corresponding scales [70]. Specifically, an error at a coarser scale will be paid the same cost as an error at a finer scale whereas it is desirable to have a higher cost for errors at coarser levels because they may generally

lead to the misclassification of groups of pixels at level 0 (e.g., one pixel at the root corresponds to 4^R pixels at the finest scale, when considering a 2-by-2 hierarchical grid).

The *marginal posterior mode (MPM)* rule is based on a criterion function that aims at segmentation accuracy and allows errors on distinct scales to be penalized differently. The cost function of MPM is:

$$C_{MPM}(\mathbf{X}, \mathbf{X}') = \sum_{s \in S} \left[1 - \delta\left(x_s, x'_s\right)\right], \tag{7.25}$$

which is related to the number of sites in which two label configurations differ. The MPM criterion penalizes errors according to their number, consequently to the scale at which they occur. The related Bayesian estimator is given by:

$$\forall s \in S, \quad \hat{x}_s = \arg\max_{x_s \in \Lambda} P(x_s|\mathbf{Y}), \tag{7.26}$$

which produces the configuration that maximizes at each site s the *a posteriori* marginal distribution of x_s conditioned to all observations $\mathbf{Y}$.

Furthermore, as shown in [19], MPM adapts well to the quad-tree topology. Indeed, because the tree is acyclic, the labels are estimated recursively by MPM through a forward-backward algorithm similar to the classical Baum and Weiss algorithm for Markov chains [6].

7.3.2 Multitemporal MPM Inference

In this section we extend the classical MPM to the described multitemporal hierarchical structure with the goal of supporting the joint classification of input images coming at multiple spatial resolutions on different times. As will be exemplified in Sect. 7.3.5 and has been mentioned in Sect. 7.1.2, multisensor input data can also be supported by resorting to copula functions.

The posterior marginal $P(x_s|\mathbf{Y})$ of the label of each spatio-temporal node s can be recursively expressed as a function of the posterior marginal $P(x_{s^-}|\mathbf{Y})$ of the parent node s^- in the corresponding quad-tree and the posterior marginal $P(x_{s^=}|\mathbf{Y})$ of the parent node $s^=$ in the quad-tree associated with the previous date, to characterize the temporal correlations associated, at different scales, with distinct images in the input time series. Indeed, the posterior marginal can be written as follows:

$$\underbrace{P(x_s|\mathbf{Y})}_{} = \sum_{x_{s^-}, x_{s^=} \in \Lambda} P(x_s|x_{s^-}, x_{s^=}, \mathbf{Y}_s) \cdot \underbrace{P(x_{s^-}|\mathbf{Y})}_{} \cdot \underbrace{P(x_{s^=}|\mathbf{Y})}_{}, \quad (7.27)$$

where underbraces denote the marginal posteriors of interest to the MPM and $\mathbf{Y}_s = \{\mathbf{y}_{s'}\}_{s' \in d(s)}$ collects the observations of all descendants of site s.

This equation involves two conditional independence assumptions:

A1 The label x_s, given the labels of its parents x_{s^-} and $x_{s^=}$ on the same and the previous dates, depends only on the observations $\mathbf{Y}_s$ of site s and its descendants but not on those of the other sites, i.e., $P(x_s|\mathbf{Y}, x_{s^-}, x_{s^=}) = P(x_s|\mathbf{Y}_s, x_{s^-}, x_{s^=})$;

A2 Given the observations, the label of the parent s^- of a site s on the same date is independent on the label of the parent $s^=$ on the previous date, i.e., $P(x_{s^-}|x_{s^=}, \mathbf{Y}) = P(x_{s^-}|\mathbf{Y})$.

These assumptions are analogous to the conditional independence assumptions that are commonly accepted in hierarchical and single-scale MRF-based image analysis. These assumptions lead to the expression (7.27) for the posterior marginals.

This formulation allows calculating recursively the posterior marginal $P(x_s|\mathbf{Y})$ at each spatio-temporal node s. Indeed, using arguments similar to [75], the following statement can easily be derived:

$$P(x_s|x_{s^-}, x_{s^=}, \mathbf{Y}_s) \propto P(x_s, x_{s^-}, x_{s^=}|\mathbf{Y}_s) = \frac{P(x_s|x_{s^-}, x_{s^=}) P(x_{s^-}|x_{s^=}) P(x_{s^=}) P(x_s|\mathbf{Y}_s)}{P(x_s)}, \quad (7.28)$$

Note that this statement holds under the following additional assumption:

A3 The distribution of the labels s^- and $s^=$ of the parents of a site s are independent on the observations $\mathbf{Y}_s$ of the descendants of s, when conditioned to the label x_s of s, i.e., $P(x_{s^-}, x_{s^=}|x_s, \mathbf{Y}_s) = P(x_{s^-}, x_{s^=}|x_s)$.

In (7.28), the first factor on the right hand side $P(x_s|x_{s^-}, x_{s^=})$ corresponds to the child-parent transition probability; $P(x_s)$ is the prior probability; $P(x_{s^-}|x_{s^=})$ is the temporal transition probability in the same scale; and $P(x_s|\mathbf{Y}_s)$ is the partial posterior marginal probability. To compute these probabilities, we benefit from the hierarchical structure defined above and use three recursive passes on the quad-tree, including one "bottom-up" and two "top-down" passes described in the following. For brevity, only the steps associated with a pair of images acquired on two different times ($t = 0$ and $t = 1$) are explained, see Fig. 7.7. The recursive extension to more than two acquisition times is straightforward.

I. Time t = 0: single-date MPM. According to the cascade approach, first, classification is performed at time $t = 0$ using a single-date MPM as in [75], where the segmentation is obtained recursively over scales through a top-down and a bottom-up stages. Details of this single-date formulation can be found in [75]. We only

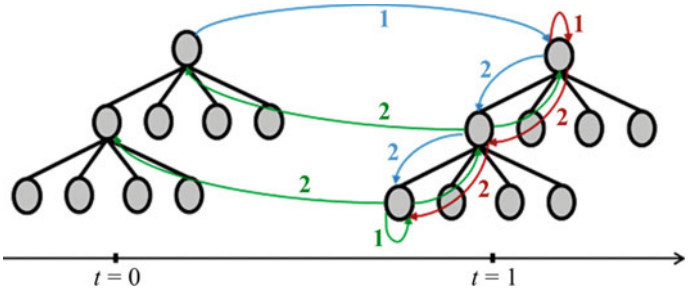


Fig. 7.7 Multitemporal recursive formulation of the MPM criterion on two quad-trees of depth $R = 2$ corresponding to two acquisition dates. *Blue*, *green*, and *red arrows* indicate the calculations performed by the first *top–down*, the *bottom–up*, and the second *top–down* passes. For each pass, numbers 1 and 2 indicate initialization and recursive computation, respectively

recall that the process is initialized by predefining the pixelwise prior probability distribution on the root of the corresponding quad-tree, i.e., $P(x_s)$, $s \in S_0^R$. This initialization is required to begin a top-down recursion and compute the priors in all levels of the quad-tree at time 0. A simple initialization strategy is to use a uniform prior distribution on Λ. Here, to incorporate spatial contextual information and mitigate possible blocky artifacts [115], a case-specific initialization strategy is applied that makes use of a spatial MRF model: a neighborhood system is defined on the lattice S_0^R in the root at time 0, and for each pixel $s \in S_0^R$, the unconditional prior $P(x_s)$ is replaced by the local conditional prior $P(x_s|x'_s, s' \sim s, s' \in S_0^R)$, where $s \sim s'$ denotes that the sites s and s' are neighbors. This choice generally provides a biased prior-probability estimate but favors spatial adaptivity, a desired property when working with high resolution images in which spatial details are common.

The well-known Potts MRF model, which favors the same labelling in homogeneous image regions, is used [70], i.e.:

$$p\left(x_s|x_{s'}, s' \sim s, s' \in S_0^R\right) \propto \exp\left[-\beta \sum_{s \sim s'} \delta\left(x_s, x_{s'}\right)\right], \tag{7.29}$$

where β is a positive spatial smoothness parameter. Several methods have been proposed to optimize the value of this parameter including the maximization of the pseudo-likelihood function over the training set [58]. As a result of single-time processing at time $t = 0$, the posterior marginal $P(x_s|\mathbf{Y})$ is known for each pixel of the corresponding quad-tree; $P(x_s|\mathbf{Y}_s)$ is also derived as a by-product ($s \in S_0^n$, $n = 0, 1, \ldots, R$), see in [75].

II. Time t = 1: first top-down pass. In the proposed method, the recursive top-down / bottom-up formulation used for the single-time case in [75] is extended to the multitemporal classification at time $t = 1$. In this case as well, first, the prior distribution on the root lattice, i.e., $P(x_s)$, $s \in S_1^R$, has to be defined to initialize a top-down pass. Following the cascade approach, we take advantage of the processing

conducted at time $t = 0$: for each pixel $s \in S_1^R$ the unconditional prior $P(x_s)$ is initialized as the posterior marginal $P(x_{s^\#}|\mathbf{Y}_{s^\#})$, which corresponds to the same pixel location $s^\# \in S_0^R$ in the root lattice S_0^R at $t = 0$ (blue arrow labeled '1' in Fig. 7.7) and has been computed as a by-product of the single-date MPM application at time $t = 0$.

After initializing the prior in the root, a top-down pass (blue arrow labeled '2' in Fig. 7.7) is performed for each finer level $n < R$ at time $t = 1$. The prior distribution is derived as a function of the prior distribution at the parent level and of the transition probabilities from the parent to the current level ($s \in S_1^n, n = 0, 1, \ldots, R-1$):

$$P(x_s) = \sum_{x_{s^-} \in \Lambda} P(x_s|x_{s^-})P(x_{s^-}). \tag{7.30}$$

This derivation favours an identical parent-child labelling and models the statistical interactions between consecutive levels of the quad-tree. We model the transition probability in the form introduced by Bouman et al. [19], i.e. ($s \in S_1^n, n = 0, 1 \cdots, R-1$):

$$P(x_s|x_{s^-}) = \begin{cases} \theta, & x_s = x_{s^-} \\ \frac{1-\theta}{M-1}, & x_s \neq x_{s^-} \end{cases}, \tag{7.31}$$

where θ is a parameter ranging in $[1/M, 1]$. As a result of the first top-down pass, the prior $P(x_s)$ is derived for each pixel $s \in S_1^n$, with $n = 0, 1, \ldots, R$ at time $t = 1$.

III. Time t = 1: **bottom-up pass**. A bottom-up pass recursion is then performed to estimate the joint probabilities $P(x_s, x_{s^-}, x_{s^=}|\mathbf{Y}_s)$ starting from the leaves of the quad-tree at $t = 1$ and proceeding until the root is reached based on (7.28).

In addition to priors, which have been computed in the previous top-down pass, three sets of probabilities are required to compute this factorization: (i) the set of temporal transition probabilities at the same scale $P(x_{s^-}|x_{s^=})$; (ii) the child-parent transition probability $P(x_s|x_{s^-}, x_{s^=})$; and (iii) the partial posterior marginals $P(x_s|\mathbf{Y}_s)$. We will present the derivations for the quantities (i) and (ii) in the next subsection.

The partial posterior marginals $P(x_s|\mathbf{Y}_s)$ are proven to satisfy ($s \in S_1^n, n = 1, 2, \cdots, R$) [75]:

$$P(x_s|\mathbf{Y}_s) \propto p(\mathbf{y}_s|x_s)P(x_s) \prod_{s' \in s^+} \sum_{x_{s'} \in \Lambda} \frac{P(x_{s'}|\mathbf{Y}_{s'})P(x_{s'}|x_s)}{P(x_{s'})}, \tag{7.32}$$

which recursively relates the value on a node to those on the offspring nodes.

The bottom-up pass proceeds recursively starting at the leaves of the quad-tree with $P(x_s|\mathbf{y}_s) \propto p(\mathbf{y}_s|x_s)P(x_s)$ (green arrow labeled '1' in Fig. 7.7) all the way until the root is reached using (7.32) (green arrow labeled '2' in Fig. 7.7). Calculation of (7.32) involves the pixelwise class-conditional PDFs $p(\mathbf{y}_s|x_s)$, whose derivation we

also present in the next subsection. As a result of the bottom-up pass, we can now compute $P(x_s, x_{s^-}, x_{s^=}|\mathbf{Y}_s)$ at each level of the quad-tree.

IV. Time t = 1: **second top-down pass**. According to (7.27), first, the posterior marginal is initialized at the root of time $t = 1$ (red arrow labeled '1' in Fig. 7.7). For this purpose, we initialize $P(x_s|\mathbf{Y})$ as $P(x_s|\mathbf{Y}_s)$ for $s \in S_1^R$, as in the usual single-date formulations of MPM [75]. Then, the posterior $P(x_s|\mathbf{Y})$ at each pixel s for all other levels at time $t = 1$ ($s \in S_1^n, n = 0, 1, \ldots, R-1$) can be easily computed recursively in a top-down pass (red arrow labeled '2' in Fig. 7.7) using (7.27).

V. Combination with MMD. At each time $t \in \{0, 1\}$, the above steps lead to the computation of the posterior marginal $P(x_s|\mathbf{Y})$ on each pixel $s \in S_t^n$, $n = 0, 1, \ldots, R$. In principle, the class label x_s that maximizes $P(x_s|\mathbf{Y})$ over Λ could be selected and assigned to s. This is often avoided in the literature of hierarchical MRFs because of its computational burden (linear with respect to the number of classes and the number of sites in all scales and times) and of possible blocky artifacts, see in [115]. As an alternate approach, here, a case-specific hybrid of the iterated conditional mode (ICM) and modified Metropolis dynamics methods for MRF energy minimization [70] is applied separately for each scale and time. This hybrid employs random sampling for the site selection and label proposal and uses a deterministic rule for the new label acceptance. We will refer to this algorithm as MMD.

In the case of the root layer of the quad-tree corresponding to each time t, MMD is used to minimize the following energy with respect to the label configuration $\mathbf{X}_t^R$:

$$U(\mathbf{X}_t^R|\mathbf{Y}) = -\sum_{s \in S_t^R} \ln P(x_s|\mathbf{Y}) - \beta \sum_{s,s' \in S_t^R, s \sim s'} \delta(x_s, x_{s'}), \tag{7.33}$$

where the first term is expressed in terms of the pixelwise posteriors computed by MPM and the second contribution is due to the Potts model on the root of the tree. MMD is iterative and is initialized with a randomly generated configuration of the label field $\mathbf{X}_t^R$. At each iteration, it randomly draws one pixel $s \in S_t^R$ and a candidate label for s using a uniform distribution: if this label yields a decrease in $U(\mathbf{X}_t^R|\mathbf{Y})$, then it is assigned to s; otherwise, it is discarded [70].

In the case of each other layer $n = 0, 1, \ldots, R-1$, no Potts model is used and MMD is applied to minimize:

$$U\left(\mathbf{X}_t^n|\mathbf{Y}\right) = -\sum_{s \in S_t^n} \ln P(x_s|\mathbf{Y}), \tag{7.34}$$

i.e., in this case, MMD is equivalent to iteratively selecting a random subset of pixels for which random replacements in class membership are attempted. In all cases, the iterative procedure of MMD is repeated until the difference in energy on consecutive iterations goes below a predefined threshold (set to $\Delta U_{min} = 10^{-4}$ in the experiments).

In the case of the root layer, the solutions obtained using MMD and maximizing $P(x_s|\mathbf{Y})$ directly intrinsically differ because the former takes into account spatial

context through the Potts model while the latter does not. In the case of the other layers, MMD acts as a randomized and computationally faster version of the maximization of $P(x_s|\mathbf{Y})$ in every pixel.

7.3.3 Transition Probabilities

The transition probabilities between consecutive scales and dates determine the properties of the hierarchical MRF because they formalize the causality of the statistical interactions involved. In the proposed method, two types of probabilities involve time. The first is the set of temporal transition probabilities at the same scale $P(x_{s^-}|x_{s^=})$, which are estimated using a specific formulation of the expectation-maximization (EM) algorithm [47]. An iterative fixed-point EM-like algorithm is performed to estimate the prior joint probabilities $P(x_{s^-}, x_{s^=})$ for each scale n, and the temporal transition probabilities are then derived [22]. The probabilities $J_{\ell m} = P(x_{s^-} = \ell, x_{s^=} = m)$ $(\ell, m \in \Lambda = \{0, 1, \ldots, M-1\})$ are regarded as the elements of an $M \times M$ matrix J, which is computed by maximizing the following pseudo-likelihood $(n = 0, 1, \ldots, R)$:

$$L(J) = \prod_{s \in S_1^n} \sum_{\ell=0}^{M-1} \sum_{m=0}^{M-1} J_{\ell m}\, p(\mathbf{y}_{s^-}, \mathbf{y}_{s^=} | x_{s^-} = \ell, x_{s^=} = m). \tag{7.35}$$

The recursive equation to be used to maximize (7.35) writes as:

$$J_{\ell m}^{\text{new}} \propto \sum_{s \in S_1^n} \frac{J_{\ell m}\, p(\mathbf{y}_{s^-}|x_{s^-} = \ell)\, p(\mathbf{y}_{s^=}|x_{s^=} = m)}{\sum_{h,k=0}^{M-1} J_{hk}\, p(\mathbf{y}_{s^-}|x_{s^-} = h)\, p(\mathbf{y}_{s^=}|x_{s^=} = k)}, \tag{7.36}$$

and is initialized flatly with $J_{\ell m} = 1/M^2$ for all $\ell, m = 0, 1, \ldots, M-1$.

The second type of transition probabilities that involve time is the child-parent transition probability $P(x_s|x_{s^-}, x_{s^=})$. To our best knowledge, a case-specific formulation of EM is not available for inter-scale transition probabilities. However, parametrically modelling these probabilities have demonstrated an effective choice in the case of single-date classification [75]. We extend here the model in [19], which favours the identity between the children and parents (in the current and previous dates), all other transitions being unlikely:

$$P(x_s|x_{s^-}, x_{s^=}) = \begin{cases} \theta, & x_s = x_{s^-} = x_{s^=} \\ \varphi, & (x_s = x_{s^-} \text{ or } x_s = x_{s^=}) \text{ and } x_{s^-} \neq x_{s^=} \\ \frac{1-\theta}{M-1}, & x_s \neq x_{s^-} \text{ and } x_s \neq x_{s^=} \text{ and } x_{s^-} = x_{s^=} \\ \frac{1-2\varphi}{M-2}, & x_s \neq x_{s^-} \text{ and } x_s \neq x_{s^=} \text{ and } x_{s^-} \neq x_{s^=} \end{cases} \tag{7.37}$$

with the parameters $\theta > 1/M$ and $1/M < \varphi < 1/2$. Here, θ has the same meaning as in (7.31), and the same value is used in both transition probabilities.

7.3.4 Pixelwise Class-Conditional PDFs

Given a training set for each input date, for each class m, scale n and acquisition time t we model the corresponding class-conditional marginal PDF $p(\mathbf{y}_s|x_s = m)$ using finite mixtures of independent distributions:

$$p(\mathbf{y}_s|x_s = m) = \sum_{k=1}^{K^{mnt}} \pi_k^{mnt} F_k^{mnt}\left(\mathbf{y}_s|\boldsymbol{\psi}_k^{mnt}\right), \qquad \forall s \in S_t^n, \tag{7.38}$$

where π_k^{mnt} are the mixing proportions, $\boldsymbol{\psi}_k^{mnt}$ is the vector of the parameters of the kth PDF mixture component of class m at scale level n and time t, and $F_k^{mnt}(\cdot)$ is the corresponding parametric family ($n = 0, 1, \ldots, R$; $m = 0, 1, \ldots, M-1$; $t = 0, 1$).

When the data at scale n and time t are multispectral or SAR, each class-conditional marginal PDF $p(\mathbf{y}_s|x_s = m)$ is modeled by a multivariate Gaussian mixture [76] or a generalized Gamma mixture [73], respectively. These assumptions, especially when combined with finite mixtures, are widely accepted for these types of data. The use of finite mixtures instead of single PDFs offers the possibility to consider heterogeneous PDFs, usually reflecting the contributions of different materials present in each class. This class heterogeneity is relevant when we address VHR images. The parameters $\boldsymbol{\psi}_k^{mnt}$ and π_k^{mnt} are estimated through the stochastic expectation maximization (SEM) algorithm [32], which is an iterative stochastic parameter estimation algorithm developed for problems characterized by data incompleteness and approaching, under suitable assumptions, maximum likelihood estimates. For each scale and time, SEM is separately applied to the training samples of each class to estimate the related parameters. We note that SEM also automatically estimates the number of mixture components K^{mnt} [74]. Only the maximum number of such components has to be predefined (and, in our experiments, was fixed to 10).

7.3.5 Experimental Study

In this section, we present several results of the developed hierarchical classifier on two datasets:

D3 A three-date time series of panchromatic and multispectral Pléiades images acquired over Port-au-Prince (Haiti) in 2011, 2012, and 2013, provided by the French Space Agency (CNES).

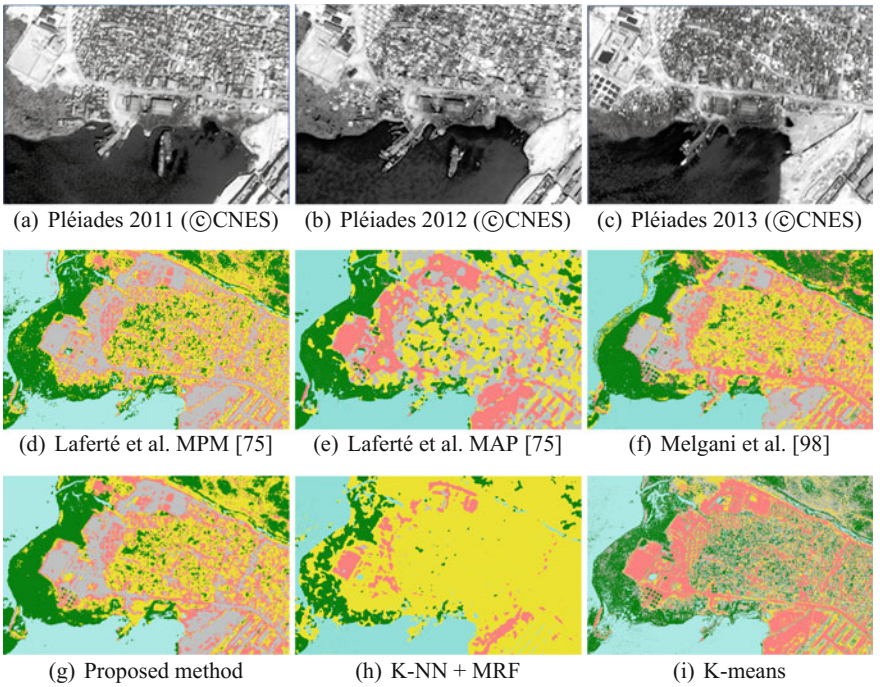
D4 A multisensor optical-SAR dataset consisting of one pansharpened GeoEye image provided by GeoEye Inc. and Google crisis response, and one HH-polarised COSMO-SkyMed acquisition of Port-au-Prince (Haiti) in 2010 provided by the Italian Space Agency (ASI).

More experimental analysis with these and other datasets is reported in [64, 149]. Five land cover classes have been considered for both data sets: *urban*, *water*, *vegetation*, *bare soil*, and *containers*. These classes represent semantically high level land covers. The spatially disjoint training and test samples have been annotated manually inside homogeneous areas. In the case of the Pléiades images, the finest resolution of the multiresolution pyramid (level 0) was set equal to the finest resolution of the panchromatic data (i.e., 0.5 m). Co-registered multispectral images (at 2 m) were integrated in level 2 of the pyramid. To fill level 1, a wavelet decomposition of the panchromatic image was used. As a preliminary experiment, the experimental analysis of the appropriateness of various wavelet transforms have been performed. It has demonstrated Daubechies 10 wavelets as the most appropriate, see in [64].

The proposed method depends on four parameters which have the following values: $\beta = 0.8$ in Eqs. (7.29) and (7.33), $\theta = 0.85$ in Eqs. (7.31) and (7.37), $\varphi = 0.48$ in Eq. (7.37), and $R = 2$. The value of β was automatically optimized by applying the pseudo-likelihood method to the training samples [70]. For (7.37) to define a probability distribution, θ and φ can take values in [0, 1] and [0, 0.5], respectively, in the case of $M = 5$ classes. The experimental study performed in [64] reported that the overall accuracy of the method grows with the larger values of these two parameters until approximately $\theta = 0.8$ and $\varphi = 0.4$, where the accuracy reaches a plateau.

In this section, we present the classification maps and discuss the corresponding classification accuracies that were obtained on the test set. Figure 7.8 and Table 7.3 report the results obtained on the multitemporal optical dataset D3, and Fig. 7.10 and Table 7.4 present those obtained on the multisensor dataset D4. The reported computation times refer to a C++ implementation on an Intel i7 quad-core (2.40 GHz) 8-GB-RAM 64-bit Linux system. The analysis of the classification maps has suggested that the proposed hierarchical method leads to accurate results. To allow for a better evaluation of the obtained results we report several comparisons with benchmark methods exploiting multiresolution, multisensor or multidate techniques.

Multitemporal D3 dataset. The results of the proposed technique were compared to the separate hierarchical classification results obtained at individual dates using the multiresolution single-time method in [75]. This comparison suggests the higher effectiveness of the proposed hierarchical model in fusing the temporal, spatial, and multiresolution information associated with the input data, see Table 7.3. In practice, the use of one quad-tree structure with the MPM criterion yields "blocky" segmentation, see Fig. 7.9. This phenomenon can be explained by the fact that two neighboring sites at a given scale may not have the same parent. In this case, a boundary appears more easily than when they are linked by a parent node. These blockiness is avoided by the introduction of causalilty over both time and scale.



(a) Pléiades 2011 (©CNES) (b) Pléiades 2012 (©CNES) (c) Pléiades 2013 (©CNES)

(d) Laferté et al. MPM [75] (e) Laferté et al. MAP [75] (f) Melgani et al. [98]

(g) Proposed method (h) K-NN + MRF (i) K-means

Fig. 7.8 Classification maps obtained on a temporal series of multiresolution Pléiades images, (© CNES distribution Airbus DS; displayed after histogram equalization). Classes: urban, water, vegetation, bare soil, and containers.

Table 7.3 Classification accuracies on the multitemporal optical Port-au-Prince dataset

Method	Urban (%)	Water (%)	Vegetation (%)	Bare soil (%)	Containers (%)	Overall (%)	Time (%)
Proposed method	81.62	100	90.69	92.82	62.82	85.59	480 s
Laferté et al. MPM	77.45	88.62	72.59	86.02	57.02	76.34	160 s
Laferté et al. MAP	56.14	100	81.90	87.02	73.21	79.65	220 s
Melgani et al.	80.63	100	86.33	87.61	69.61	84.83	$\approx$ 1h
K-NN + MRF	96.84	92.42	47.15	71.83	16.75	64.99	90 s
K-means	12.37	98.63	59.18	91.66	29.42	58.25	20 s

Table 7.4 Classification accuracies on the multisensor optical-SAR Port-au-Prince dataset

Method	Water (%)	Urban (%)	Vegetation (%)	Bare soil (%)	Containers (%)	Overall (%)
Proposed SAR+opt.	100	75.24	87.16	98.89	49.31	82.12
Proposed opt.	100	67.12	86.89	98.83	41.90	78.95
Single-scale MRF	100	100	81.42	99.94	59.62	88.20
Storvik et al. [136]	99.95	79.32	90.81	96.22	37.25	79.44
GML [14] + MRF	99.99	84.51	93.24	99.68	58.64	86.28
K-means	100	39.01	91.19	82.32	6.54	57.30

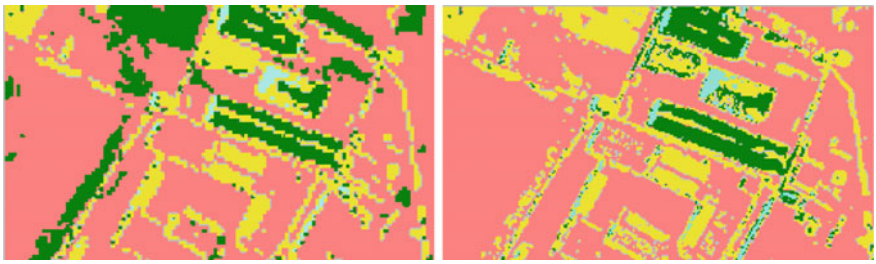
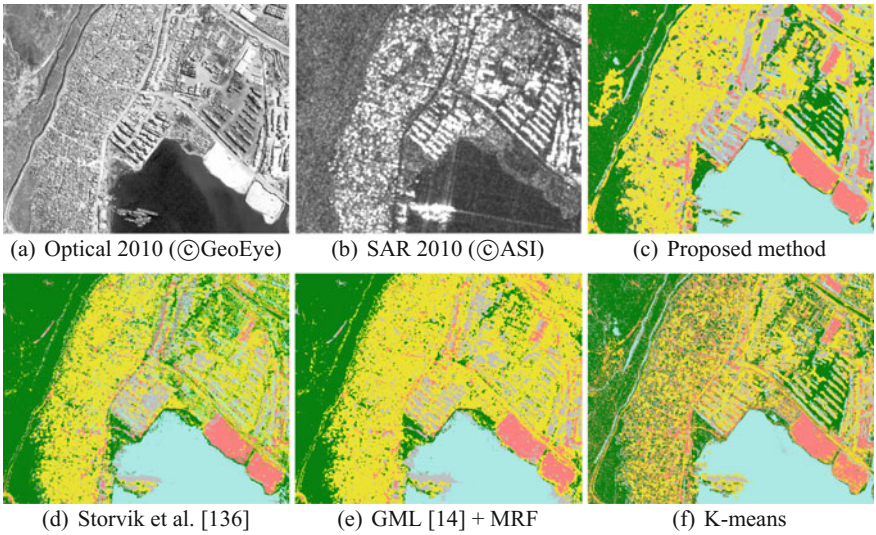


Fig. 7.9 Zoom-ins of two classification maps from Fig. 7.8: blocky artefacts obtained using the Laferté et al. [75] MPM formulation (*left*), and reduction of these artefacts using the proposed method (*right*)

As expected, the MAP criterion was strikingly less efficient when applied to the considered hierarchical structure because errors were propagated from the root to the leaves and led to severe misclassification, especially regarding the classes that most strongly overlap in the feature space (e.g., "urban" and "containers").

The proposed classifier was further compared to the multitemporal single-resolution MRF-based method in [98]. It uses the mutual approach and consists in performing a bidirectional exchange of the temporal information between the (non-hierarchical) single-time MRF models associated with consecutive images in the sequence. The accuracy comparison shows a better exploitation of the spatio-temporal information by the proposed approach. More generally, the mutual approach reduces the risk of propagating the classification error between consecutive dates, while the use of the hierarchical schema provided more accurate and faster (due to its non-iterative nature) classification maps. Next, the classical k-nearest neighbours (k-NN) was used as a benchmark non-parametric classifier with $k = 30$ estimated by cross-validation. It is non-contextual, so to perform a fair comparison between



(a) Optical 2010 (©GeoEye) (b) SAR 2010 (©ASI) (c) Proposed method

(d) Storvik et al. [136] (e) GML [14] + MRF (f) K-means

Fig. 7.10 Classification maps obtained on a multisensor GeoEye 2010 (optical, © GeoEye) + COSMO-SkyMed 2010 (SAR, © ASI) Port-au-Prince dataset (displayed after histogram stretching and equalization). Same color legend as in Fig. 7.8

the proposed method and a spatial-contextual technique, it was combined with an isotropic Potts MRF model. It is immediate that this single-scale MRF leads to severe spatial oversmoothing. Finally, a further comparison was performed with K-means initialized with $K = 5$. As expected, the clusters obtained by K-means do not match well with the thematic classes of the considered problem.

Multisensor D4 dataset. The GeoEye image resolution comes at the 0.5 m resolution, and the COSMO-SkyMed image with 2.5 m pixel spacing. To fit with the dyadic decomposition imposed by the quad-tree, we slightly resampled the optical image to obtain the $0.625 = 2.5/4$ m resolution that was put at the finest resolution of the pyramid. We apply copula functions to model dependence between optical and SAR data, whose marginals are estimated using mixtures of Gaussian and generalized Gamma distributions, see [149]. Since all the data correspond to the same date, the images are integrated into a single pyramid comprised of three levels.

The proposed hierarchical MPM-based method, see Fig. 7.10, leads to a detailed classification with an adequate level of classification map regularity. The main source of misclassification is the "container" class, where the asphalt is erroneously classified as vegetation. In Table 7.4 we compare numerically the results obtained with the proposed hierarchical method when considering either only optical, or both SAR and optical images. We observe an improvement related to the combination of the two images, in particular in the urban areas for which the SAR acquisition represents a significant source of discriminative information. More specifically, we have

observed that the optical image has a relevant effect in the "bare soil" discrimination, and the SAR acquisition is particularly helpful to identify the "containers".

Numerical results suggest that the single-scale MRF-based method leads to the highest accuracy. However, this comes at the price of severe oversmoothing. This is a result of selecting the β parameter maximizing the classification accuracies. Consequently, the localization of the ground truth within homogeneous regions brought to an excessive oversmoothing of the edges. The second best result is demonstrated by the GML+MRF method. The visual inspection suggests that the resulting classification map is highly fragmented, especially as compared with the map reported by the proposed method. A closer look into the "container" area reveals a better structured and (visually) more accurate classification with a spatially more precise characterization of the geometrical structure reported by the proposed approach. We then compare the proposed method with the single-scale multi-sensor approach in [136]; in this case, the SAR image is upscaled and the likelihood term is constructed by merging the two generalized Gamma marginals into a meta-Gaussian distribution. The classification is obtained by maximum likelihood with an isotropic 3-by-3 Potts-model MRF with optimization by ICM with $\beta = 1.70$. This method, as well as K-means (grouped version, with $K = 10$ initialization) demonstrate a considerably more noisy output and perform worse with the "containers" class.

7.4 Conclusions

The present chapter was the second of a pair of chapters dedicated to remote sensing data fusion. The methodological issues associated with multiple information sources within the processing pipeline for image classification have been addressed. With regard to the feature extraction stage, the problem of the computation of descriptors associated with distinct spatial scales have been discussed, generalized to the framework of multilevel feature extraction, and formalized through advanced methods rooted in the theories of Mathematical Morphology and attribute profiles. Concerning the classification stage, the task of the joint supervised classification of multisensor, multiresolution, multiscale, and/or multitemporal images has been addressed by recalling the main methodological strategies and by focusing on advanced models based on hierarchical Markovian formalizations on quad-trees.

In both cases, the examples of experimental results have suggested the effectiveness of the morphological and Markovian approaches to the extraction and fusion of multilevel and multisource information for classification purposes. Following up on the conclusions of the previous chapter, these results further confirm that current mathematical models for remote sensing data fusion, stemming from the fields of pattern recognition, stochastic modelling, and graph theory, represent powerful, flexible, and computationally efficient tools that allow taking benefit from the variety of remote sensing data sources available nowadays.

Acknowledgements This work was partly supported by the French Space Agency (*Centre National d'Etudes Spatiales*, CNES) through contract no. 8361. The authors would like to thank CNES, the Italian Space Agency (ASI), and GeoEye Inc. and Google Crisis Response for providing the Pléiades, COSMO-SkyMed, and GeoEye imagery used for experiments.

References

1. Akcay, H.G., Aksoy, S.: Automatic detection of geospatial objects using multiple hierarchical segmentations. IEEE Trans. Geosci. Remote Sens. **46**(7), 2097–2111 (2008)
2. Alonso-Gonzalez, A., Valero, S., Chanussot, J., Lopez-Martinez, C., Salembier, P.: Processing multidimensional SAR and hyperspectral images with binary partition tree. Proc. IEEE **101**(3), 723–747 (2013)
3. Amici, G., Dell'Acqua, F., Gamba, P., Pulina, G.: A comparison of fuzzy and neuro-fuzzy data fusion for flooded area mapping using SAR images. Int. J. Remote Sens. **25**(20), 4425–4430 (2004)
4. Atkinson, P.M., Aplin, P.: Spatial variation in land cover and choice of spatial resolution for remote sensing. Int. J. Remote Sens. **25**(18), 3687–3702 (2004)
5. Bakos, K., Gamba, P.: Hierarchical hybrid decision tree fusion of multiple hyperspectral data processing chains. IEEE Trans. Geosci. Remote Sens. **49**(1), 388–394 (2011)
6. Baum, L., Petrie, T., Soules, G., Weiss, N.: A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. Ann. Math. Stat. pp. 164–171 (1970)
7. Benediktsson, J.A.: Classification of multisource and hyperspectral data based on decision fusion. IEEE Trans. Geosci. Remote Sens. **37**(3), 1367–1377 (1999)
8. Benediktsson, J.A., Bruzzone, L., Chanussot, J., Dalla Mura, M., Salembier, P., Valero, S.: Hierarchical analysis of remote sensing data: morphological attribute profiles and binary partition trees. In: Mathematical Morphology and Its Applications to Image and Signal Processing, vol. 6671 LNCS, pp. 306–319. Springer, Berlin (2011)
9. Benediktsson, J.A., Palmason, J.A., Sveinsson, J.R.: Classification of hyperspectral data From urban areas based on extended morphological profiles. IEEE Trans. Geosci. Remote Sens. **43**(3), 480–491 (2005)
10. Benediktsson, J.A., Pesaresi, M., Arnason, K.: Classification and feature extraction for remote sensing images from urban areas based on morphological transformations. IEEE Trans. Geosci. Remote Sens. **41**(9), 1940–1949 (2003)
11. Bernabe, S., Marpu, P.R., Plaza, A., Mura, M.D., Benediktsson, J.A.: Spectral-spatial classification of multispectral images using kernel feature space representation. IEEE Geosci. Remote Sens. Lett. **11**(1), 288–292 (2014)
12. Beucher, S., Meyer, F.: The morphological approach to segmentation: the watershed transformation. Opt. Eng. **34**, 433–481 (1993)
13. Bigdeli, B., Samadzadegan, F., Reinartz, P.: A decision fusion method based on multiple support vector machine system for fusion of hyperspectral and LIDAR data. Int. J. Image Data Fusion **5**(3), 196–209 (2014)
14. Bishop, C.: Pattern Recognition And Machine Learning. Springer, Berlin (2006)
15. Blaschke, T.: Object based image analysis for remote sensing. ISPRS J. Photogramm. Remote Sens. **65**(1), 2–16 (2010)
16. Blaschke, T., Hay, G.J., Kelly, M., Lang, S., Hofmann, P., Addink, E., Queiroz Feitosa, R., van der Meer, F., van der Werff, H., van Coillie, F., Tiede, D.: Geographic object-based image analysis - towards a new paradigm. ISPRS J. Photogramm. Remote Sens. **87**, 180–191 (2014)
17. Bogdanov, A.: Neuroinspired architecture for robust classifier fusion of multisensor imagery. IEEE Trans. Geosci. Remote Sens. **46**(5), 1467–1487 (2008)

18. Boudaren, M.E.Y., An, L., Pieczynski, W.: Dempster-Shafer fusion of evidential pairwise Markov fields. Int. J. Approx. Reason. **74**, 13–29 (2016)
19. Bouman, C.A., Shapiro, M.: A multiscale random field model for Bayesian image segmentation. IEEE Trans. Image Process. **3**(2), 162–177 (1994)
20. Breen, E.J., Jones, R.: Attribute openings, thinnings, and granulometries. Comput. Vis. Image Und. **64**(3), 377–389 (1996)
21. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)
22. Bruzzone, L., Prieto, D.F., Serpico, S.B.: A neural-statistical approach to multitemporal and multisource remote-sensing image classification. IEEE Trans. Geosci. Remote Sens. **37**(3), 1350–1359 (1999)
23. Burnett, C., Blaschke, T.: A multi-scale segmentation/object relationship modelling methodology for landscape analysis. Ecol. Model. **168**(3), 233–249 (2003)
24. Camps-Valls, G., Gomez-Chova, L., Munoz-Mari, J., Rojo-Alvarez, J., Martinez-Ramon, M.: Kernel-based framework for multitemporal and multisource remote sensing data classification and change detection. IEEE Trans. Geosci. Remote Sens. **46**(6), 1822–1835 (2008)
25. Camps-Valls, G., Tuia, D., Bruzzone, L., Benediktsson, J.A.: Advances in hyperspectral image classification: Earth monitoring with statistical learning methods. IEEE Signal Process. Mag. **31**(1), 45–54 (2014)
26. Carlinet, E., Géraud, T.: A comparative review of component tree computation algorithms. IEEE Trans. Image Process. **23**(9), 3885–3895 (2014)
27. Caselles, V., Coll, B., Morel, J.M.: Topographic maps and local contrast changes in natural images. Int. J. Comput. Vision **33**(1), 5–27 (1999)
28. Caselles, V., Monasse, P.: Geometric Description Of Images As Topographic Maps, 1st edn. Springer, Berlin (1984)
29. Cavallaro, G., Falco, N., Dalla Mura, M., and J. A. Benediktsson.: Automatic Attribute Profiles. IEEE Trans. Image Process. **26**(4), 1859–1872 (Apr 2017)
30. Cavallaro, G., Dalla Mura, M., Benediktsson, J. A., Bruzzone, L.: Extended self-dual attribute profiles for the classification of hyperspectral images. IEEE Geosci. Remote Sens. Lett. **99**(8), 1–5 (2015)
31. Ceamanos, X., Waske, B., Benediktsson, J.A., Chanussot, J., Fauvel, M., Sveinsson, J.: A classifier ensemble based on fusion of support vector machines for classifying hyperspectral data. Int. J. Image Data Fusion **1**(4), 293–307 (2010)
32. Celeux, G., Chauveau, D., Diebolt, J.: Stochastic versions of the EM algorithm: an experimental study in the mixture case. J. Stat. Comput. Sim. **55**(4), 287–314 (1996)
33. Chanussot, J., Mauris, G., Lambert, P.: Fuzzy fusion techniques for linear features detection in multitemporal SAR images. IEEE Trans. Geosci. Remote Sens. **37**(3 I), 1292–1305 (1999)
34. Coburn, C.A., Roberts, A.C.B.: A multiscale texture analysis procedure for improved forest stand classification. Int. J. Remote Sens. **25**(20), 4287–4308 (2004)
35. Crozet, S., Géraud, T.: A first parallel algorithm to compute the morphological tree of shapes of nD Images. In: Proceedings of the IEEE International Conference on Image Processing, pp. 2933–2937 (2014)
36. Dalla Mura, M., Benediktsson, J.A., Bruzzone, L.: Classification of hyperspectral images with extended attribute profiles and feature extraction techniques. In: Proceedings of the IEEE International Geoscience and Remote Sensing Symposium, pp. 76–79 (2010)
37. Dalla Mura, M., Benediktsson, J.A., Bruzzone, L.: Self-dual attribute profiles for the analysis of remote sensing images. In: Mathematical Morphology and Its Applications to Image and Signal Processing, pp. 320–330. Springer, Berlin (2011)
38. Dalla Mura, M., Benediktsson, J.A., Chanussot, J., Bruzzone, L.: The evolution of the morphological profile: From panchromatic to hyperspectral images. In: Optical Remote Sensing: Advances in Signal Processing and Exploitation Techniques, pp. 123–146. Springer, Berlin (2011)
39. Dalla Mura, M., Benediktsson, J.A., Waske, B., Bruzzone, L.: Extended profiles with morphological attribute filters for the analysis of hyperspectral data. Int. J. Remote Sens. **31**(22), 5975–5991 (2010)

40. Dalla Mura, M., Benediktsson, J.A., Waske, B., Bruzzone, L.: Morphological attribute profiles for the analysis of very high resolution images. IEEE Trans. Geosci. Remote Sens. **48**(10), 3747–3762 (2010)
41. Dalla Mura, M., Villa, A., Benediktsson, J.A., Chanussot, J., Bruzzone, L.: Classification of hyperspectral images by using extended morphological attribute profiles and independent component analysis. IEEE Geosci. Remote Sens. Lett. **8**(3), 542–546 (2011)
42. Dalponte, M., Bruzzone, L., Gianelle, D.: Fusion of hyperspectral and LIDAR remote sensing data for classification of complex forest areas. IEEE Trans. Geosci. Remote Sens. **46**(5), 1416–1427 (2008)
43. Datcu, M., Melgani, F., Piardi, A., Serpico, S.B.: Multisource data classification with dependence trees. IEEE Trans. Geosci. Remote Sens. **40**(3), 609–617 (2002)
44. Dawid, A.: Applications of a general propagation algorithm for probabilistic expert systems. Stat. Comput. **2**(1), 25–36 (1992)
45. Dell'Acqua, F., Gamba, P.: Discriminating urban environments using multiscale texture and multiple SAR images. Int. J. Remote Sens. **27**(18), 3797–3812 (2006)
46. Demir, B., Bruzzone, L.: Histogram-based attribute profiles for classification of very high resolution remote sensing images. IEEE Trans. Geosci. Remote Sens. **54**(4), 2096–2107 (2016)
47. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Stat. Soc. Ser.B **39**(1), 1–38 (1977)
48. Dos Santos, J., Gosselin, P.H., Philipp-Foliguet, S., Torres, Da S.R., Falcao, A.: Multiscale classification of remote sensing images. IEEE Trans. Geosci. Remote Sens. **50**(10), 3764–3775 (2012)
49. El-melegy, M., Ahmed, S.: Neural networks in multiple classifier systems for remote-sensing image classification. Stud. Fuzziness Soft Comput. **210**, 65–94 (2007)
50. Falco, N., Benediktsson, J.A., Bruzzone, L.: Spectral and spatial classification of hyperspectral images Based on ICA and reduced morphological attribute profiles. IEEE Trans. Geosci. Remote Sens. **53**(11), 6223–6240 (2015)
51. Falco, N., Dalla Mura, M., Bovolo, F., Benediktsson, J.A., Bruzzone, L.: Change detection in VHR images based on morphological attribute profiles. IEEE Geosci. Remote Sens. Lett. **10**(3), 636–640 (2013)
52. Fauvel, M., Chanussot, J., Benediktsson, J.A.: Decision fusion for the classification of urban remote sensing images. IEEE Trans. Geosci. Remote Sens. **44**(10), 2828–2838 (2006)
53. Fauvel, M., Tarabalka, Y., Benediktsson, J.A., Chanussot, J., Tilton, J.C.: Advances in spectral-spatial classification of hyperspectral images. Proc. IEEE **101**(3), 652–675 (2013)
54. Forney, G.D.: The Viterbi algorithm. Proc. IEEE **61**(3), 268–278 (1973)
55. Foucher, S., Bénié, G.B., Boucher, J.M.: Multiscale MAP filtering of SAR images. IEEE Trans. Image Process. **10**(1), 49–60 (2001)
56. Franchi, G., Angulo, J.: Morphological principal component analysis for hyperspectral image analysis. ISPRS Int. J. Geo-Inf. **5**(6), 83 (2016)
57. Gamba, P., Houshmand, B.: An efficient neural classification chain of SAR and optical urban images. Int. J. Remote Sens. **22**(8), 1535–1553 (2001)
58. Geman, S., Geman, D.: Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Trans. Pattern Anal. Mach. Intell. **6**(6), 721–741 (1984)
59. Géraud, T., Carlinet, E., Crozet, S., Najman, L.: A quasi-linear algorithm to compute the tree of shapes of nD images. In: Mathematical Morphology and Its Applications to Signal and Image Processing, pp. 98–110. Springer, Berlin (2013)
60. Gerke, M., Xiao, J.: Fusion of airborne laser scanning point clouds and images for supervised and unsupervised scene classification. ISPRS J. Photogramm. Remote Sens. **87**, 78–92 (2014)
61. Gomez-Chova, L., Tuia, D., Moser, G., Camps-Valls, G.: Multimodal classification of remote sensing images: A review and future directions. Proc. IEEE **103**(9), 1560–1584 (2015)
62. Haralick, R.M.: Statistical and structural approaches to texture. Proc. IEEE **67**(5), 786–804 (1979)
63. Hedhli, I., Moser, G., Serpico, S.B., Zerubia, J.: New hierarchical joint classification method of SAR-optical multiresolution remote sensing data. In: Proceedings of the IEEE European Signal Processing Conference, pp. 759–763 (2015)

64. Hedhli, I., Moser, G., Serpico, S.B., Zerubia, J.: A new cascade model for the hierarchical joint classification of multitemporal and multiresolution remote sensing data. IEEE Trans. Geosci. Remote Sens. **54**(11), 6333–6348 (2016)
65. Hedhli, I., Moser, G., Zerubia, J., Serpico, S.B.: New cascade model for hierarchical joint classification of multitemporal, multiresolution and multisensor remote sensing data. In: Proceedings of the IEEE International Conference on Image Processing, pp. 5247–5251 (2014)
66. Hoberg, T., Rottensteiner, F., Feitosa, R., Heipke, C.: Conditional random fields for multitemporal and multiscale classification of optical satellite imagery. IEEE Trans. Geosci. Remote Sens. **53**(2), 659–673 (2015)
67. Jalobeanu, A., Blanc-Feraud, L., Zerubia, J.: Satellite image deblurring using complex wavelet packets. Int. J. Comput. Vision **51**(3), 205–217 (2003)
68. Jones, R.: Component trees for image filtering and segmentation. In: Proceedings of the IEEE Workshop on Nonlinear Signal and Image Processing. Mackinac Island (1997)
69. Jones, R.: Connected filtering and segmentation using component trees. Comput. Vis. Image Und. **75**(3), 215–228 (1999)
70. Kato, Z., Zerubia, J.: Markov random fields in image segmentation. Found. Trends Signal Proc. **5**(1–2), 1–155 (2012)
71. Kruskal, J.B.: On the shortest spanning subtree of a graph and the traveling salesman problem. Proc. Am. Math. Soc. **7**, 48–50 (1956)
72. Krylov, V., Moser, G., Serpico, S.B., Zerubia, J.: Supervised high-resolution dual-polarization SAR image classification by finite mixtures and copulas. IEEE J. Sel. Top. Signal Process. **5**(3), 554–566 (2011)
73. Krylov, V., Moser, G., Serpico, S.B., Zerubia, J.: On the method of logarithmic cumulants for parametric probability density function estimation. IEEE Trans. Image Process. **22**(10), 3791–3806 (2013)
74. Krylov, V., Moser, G., Serpico, S.B., Zerubia, J.: Enhanced dictionary-based SAR amplitude distribution estimation and its validation with very high-resolution data. IEEE Geosci. Remote Sens. Lett. **8**(1), 148–152 (2011)
75. Laferté, J.M., Pérez, P., Heitz, F.: Discrete Markov image modeling and inference on the quadtree. IEEE Trans. Image Process. **9**(3), 390–404 (2000)
76. Landgrebe, D.A.: Signal theory methods in multispectral remote sensing. John Wiley & Sons Inc., (2003)
77. Le Hegarat-Mascle, S., Richard, D., Ottle, C.: Multi-scale data fusion using Dempster-Shafer evidence theory. Integr. Comput.-Aid. Eng. **10**(1), 9–22 (2003)
78. Lecun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**(7553), 436–444 (2015)
79. Lee, S., Crawford, M.M.: Unsupervised multistage image classification using hierarchical clustering with a Bayesian similarity measure. IEEE Trans. Image Process. **14**(3), 312–320 (2005)
80. Lemire, D.: A better alternative to piecewise linear time series segmentation. **2007**, 545–550 (2006). arXiv:cs/0605103v8
81. Li, M., Zang, S., Zhang, B., Li, S., Wu, C.: A review of remote sensing image classification techniques: The role of spatio-contextual information. Eur. J. Remote Sens. **47**(1), 389–411 (2014)
82. Li, S.: Markov Random Field Modeling In Image Analysis, 3rd edn. Springer, Berlin (2009)
83. Liao, W., Pizurica, A., Bellens, R., Gautama, S., Philips, W.: Generalized graph-based fusion of hyperspectral and LiDAR data using morphological features. IEEE Geosci. Remote Sens. Lett. **12**(3), 552–556 (2014)
84. Liu, Z.G., Mercier, G., Dezert, J., Pan, Q.: Change detection in heterogeneous remote sensing images based on multidimensional evidential reasoning. IEEE Geosci. Remote Sens. Lett. **11**(1), 168–172 (2014)
85. Lombardo, P., Oliver, C., Pellizzeri, T., Meloni, M.: A new maximum-likelihood joint segmentation technique for multitemporal SAR and multiband optical images. IEEE Trans. Geosci. Remote Sens. **41**(11), 2500–2518 (2003)

86. Loncan, L., De Almeida, L., Bioucas-Dias, J., Briottet, X., Chanussot, J., Dobigeon, N., Fabre, S., Liao, W., Licciardi, G., Simoes, M., Tourneret, J.Y., Veganzones, M., Vivone, G., Wei, Q., Yokoya, N.: Hyperspectral pansharpening: a review. IEEE Geosci. Remote Sens. Mag. **3**(3), 27–46 (2015)
87. Luettgen, M., Karl, W., Willsky, A.: Efficient multiscale regularization with applications to the computation of optical flow. IEEE Trans. Image Process. **3**(1), 41–64 (1994)
88. Willsky, A.: Multiresolution Markov models for signal and image processing. Proc. IEEE **90**(8), 1396–1458 (2002)
89. Luus, F., Salmon, B., Van Den Bergh, F., Maharaj, B.: Multiview deep learning for land-use classification. IEEE Geosci. Remote Sens. Lett. **12**(12), 2448–2452 (2015)
90. Mahmood, Z., Thoonen, G., Scheunders, P.: Automatic threshold selection for morphological attribute profiles. In: Proceedings of the IEEE International Geoscience and Remote Sensing Symposium, pp. 4946–4949 (2012)
91. Mallat, S.: A Wavelet Tour Of Signal Processing, 3rd edn. Academic press, Dublin (2008)
92. Maragos, P.: Pattern spectrum and multiscale shape representation. IEEE Trans. Pattern Anal. Mach. Intell. **11**(7), 701–716 (1989)
93. Marceau, D.J.: The scale issue in social and natural sciences. Can. J. Remote Sens. **25**(July), 347–356 (1999)
94. Marmanis, D., Datcu, M., Esch, T., Stilla, U.: Deep learning earth observation classification using ImageNet pretrained networks. IEEE Geosci. Remote Sens. Lett. **13**(1), 105–109 (2016)
95. Marpu, P.R., Pedergnana, M., Dalla Mura, M., Benediktsson, J.A., Bruzzone, L.: Automatic generation of standard deviation attribute profiles for spectral-spatial classification of remote sensing data. IEEE Geosci. Remote Sens. Lett. **10**(2), 293–297 (2013)
96. Marpu, P.R., Pedergnana, M., Dalla Mura, M., Peeters, S., Benediktsson, J.A., Bruzzone, L.: Classification of hyperspectral data using extended attribute profiles based on supervised and unsupervised feature extraction techniques. Int. J. Image Data Fusion **3**(3), 269–298 (2012)
97. Matheron, G.: Random Sets And Integral Geometry. John Wiley & Sons, Newyork (1975)
98. Melgani, F., Serpico, S.B.: A Markov random field approach to spatio-temporal contextual image classification. IEEE Trans. Geosci. Remote Sens. **41**(11), 2478–2487 (2003)
99. Melgani, F., Serpico, S.B., Vernazza, G.: Fusion of multitemporal contextual information by neural networks for multisensor remote sensing image classification. Integr. Comput.-Aid. Eng. **10**(1), 81–90 (2003)
100. Merentitis, A., Debes, C.: Many hands make light work - on ensemble learning techniques for data fusion in remote sensing. IEEE Geosci. Remote Sens. Mag. **3**(3), 86–99 (2015)
101. Monasse, P., Guichard, F.: Fast computation of a contrast-invariant image representation. IEEE Trans. Image Process. **9**(5), 860–872 (2000)
102. Moser, G., De Giorgi, A., Serpico, S.B.: Multiresolution supervised classification of panchromatic and multispectral images by Markov random fields and graph cuts. IEEE Trans. Geosci. Remote Sens. **43**(8), 1901–1911 (2016)
103. Moser, G., Serpico, S.B., Benediktsson, J.A.: Land-cover mapping by Markov modeling of spatial-contextual information in very-high-resolution remote sensing images. Proc. IEEE **101**(3), 631–651 (2013)
104. Najman, L., Cousty, J.: A graph-based mathematical morphology reader. Pattern Recogn. Lett. **47**, 3–17 (2014)
105. Najman, L., Talbot, H.: Connected operators based on tree pruning strategies. In: Mathematical Morphology: From Theory to Applications, pp. 177–198. John Wiley & Sons, Newyork (2010)
106. Nishii, R.: A Markov random field-based approach to decision-level fusion for remote sensing image classification. IEEE Trans. Geosci. Remote Sens. **41**(10), 2316–2319 (2003)
107. Ouzounis, G.K., Pesaresi, M., Soille, P.: Differential area profiles: decomposition properties and efficient computation. IEEE Trans. Pattern Anal. Mach. Intell. **34**(8), 1533–1548 (2012)
108. Ouzounis, G.K., Soille, P.: The Alpha-tree Algorithm. Publications Office of the European Union, EUR 25500 EN (2012)
109. Ouzounis, G.K., Wilkinson, M.H.F.: Partition-induced connections and operators for pattern analysis. Pattern Recogn. **43**(10), 3193–3207 (2010)

110. Pacifici, F., Chini, M., Emery, W.J.: A neural network approach using multi-scale textural metrics from very high-resolution panchromatic imagery for urban land-use classification. Remote Sens. Environ. **113**(6), 1276–1292 (2009)
111. Palau, A., Melgani, F., Serpico, S.B.: Cell algorithms with data inflation for non-parametric classification. Pattern Recogn. Lett. **27**(7), 781–790 (2006)
112. Park, N.W., Moon, W., Chi, K.H., Kwon, B.D.: Multi-sensor data fusion for supervised land-cover classification using Bayesian and geostatistical techniques. Geosci. J. **6**(3) (2002)
113. Pedergnana, M., Marpu, P.R., Dalla Mura, M., Benediktsson, J.A., Bruzzone, L.: Classification of remote sensing optical and LiDAR data using extended attribute profiles. IEEE J. Sel. Top. Signal Process. **6**(7), 856–865 (2012)
114. Peeters, S., Marpu, P.R., Benediktsson, J.A., Dalla Mura, M.: Classification using extended morphological attribute profiles based on different feature extraction techniques. In: Proceedings of the IEEE International Geoscience and Remote Sensing Symposium, pp. 4453–4456 (2011)
115. Pérez, P., Chardin, A., Laferté, J.M.: Noniterative manipulation of discrete energy-based models for image analysis. Pattern Recogn. **33**(4), 573–586 (2000)
116. Pesaresi, M., Benediktsson, J.A.: A new approach for the morphological segmentation of high-resolution satellite imagery. IEEE Trans. Geosc. Remote Sens. **39**(2), 309–320 (2001)
117. Plaza, A., Martinez, P., Plaza, J., Perez, R.: Dimensionality reduction and classification of hyperspectral image data using sequences of extended morphological transformations. IEEE Trans. Geosci. Remote Sens. **43**(3), 466–479 (2005)
118. Poggi, G., Scarpa, G., Zerubia, J.: Supervised segmentation of remote sensing images based on a tree-structured MRF model. IEEE Trans. Geosci. Remote Sens. **54**(9), 5054–5070 (2005)
119. Pohl, C., van Genderen, J.: Remote sensing image fusion: An update in the context of digital Earth. Int. J. Digital Earth **7**(2), 158–172 (2014)
120. Quesada-Barriuso, P., Arguello, F., Heras, D.B.: Spectral-spatial classification of hyperspectral images using wavelets and extended morphological profiles. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. **7**(4), 1177–1185 (2014)
121. Ran, Y., Li, X., Lu, L., Li, Z.: Large-scale land cover mapping with the integration of multi-source information based on the Dempster-Shafer theory. Int. J. Geogr. Inf. Sci. **26**(1), 169–191 (2012)
122. Ranchin, T., Wald, L.: The wavelet transform for the analysis of remotely sensed images. Int. J. Remote Sens. **14**(3), 615–619 (1993)
123. Saeidi, V., Pradhan, B., Idrees, M., Latif, Z.: Fusion of airborne LiDAR with multispectral SPOT 5 image for enhancement of feature extraction using Dempster-Shafer theory. IEEE Trans. Geosci. Remote Sens. **52**(10), 6017–6025 (2014)
124. Salembier, P., Garrido, L.: Binary partition tree as an efficient representation for image processing, segmentation, and information retrieval. IEEE Trans. Image Process. **9**(4), 561–576 (2000)
125. Salembier, P., Oliveras, A., Garrido, L.: Antiextensive connected operators for image and sequence processing. IEEE Trans. Image Process. **7**(4), 555–570 (1998)
126. Salembier, P., Serra, J.: Flat zones filtering, connected operators, and filters by reconstruction. IEEE Trans. Image Process. **4**(8), 1153–1160 (1995)
127. Salembier, P., Wilkinson, M.: Connected operators. IEEE Signal Process. Mag. **26**(6), 136–157 (2009)
128. Scarpa, G., Gaetano, R., Haindl, M., Zerubia, J.: Hierarchical multiple Markov chain model for unsupervised texture segmentation. IEEE Trans. Image Process. **18**(8), 1830–1843 (2009)
129. Schistad Solberg, A., Taxt, T., Jain, A.: A Markov random field model for classification of multisource satellite imagery. IEEE Trans. Geosci. Remote Sens. **34**(1), 100–113 (1996)
130. Serra, J.: Image Analysis And Mathematical Morphology. Academic Press, Dublin (1982)
131. Serra, J.: Image Analysis and Mathematical Morphology. Theoretical Advances. Serra, J. (ed.), vol. 2. Journal of Microscopy (1988)
132. Simard, M., Saatchi, S.S., De Grandi, G.: The use of decision tree and multiscale texture for classification of JERS-1 SAR data over tropical forest. IEEE Trans. Geosci. Remote Sens. **38**(5), 2310–2321 (2000)

133. Soille, P.: Morphological Image Analysis: Principles And Applications, 2nd edn. Springer, Berlin (2004)
134. Song, B., Dalla Mura, M., Li, P., Plaza, A.J., Bioucas-Dias, J.M., Benediktsson, J.A., Chanussot, J.: Remotely sensed image classification using sparse representations of morphological attribute profiles. IEEE Trans. Geosci. Remote Sens. **52**(8), 5122–5136 (2014)
135. Song, Y.: A Topdown algorithm for computation of level line trees. IEEE Trans. Image Process. **16**(8), 2107–2116 (2007)
136. Storvik, B., Storvik, G., Fjortoft, R.: On the combination of multisensor data using meta-Gaussian distributions. IEEE Trans. Geosci. Remote Sens. **47**(7), 2372–2379 (2009)
137. Sutton, C., McCallum, A.: An introduction to conditional random fields. Found. Trends Mach. Learn. **4**(4), 267–373 (2011)
138. Tarabalka, Y., Benediktsson, J.A., Chanussot, J., Tilton, J.C.: Multiple spectral-spatial classification approach for hyperspectral data. IEEE Trans. Geosci. Remote Sens. **48**(11), 4122–4132 (2010)
139. Thoonen, G., Mahmood, Z., Peeters, S., Scheunders, P.: Multisource classification of color and hyperspectral images using color attribute profiles and composite decision fusion. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. **5**(2), 510–521 (2012)
140. Tilton, J.C.: Analysis of hierarchically related image segmentations. In: Proceedings of the 2003 IEEE Workshop on Advances in Techniques for Analysis of Remotely Sensed Data 00(C), 60–69 (2004)
141. Tuceryan, M., Jain, A.K.: Texture analysis. In: The Handbook of Pattern Recognition and Computer Vision, 2nd edn., pp. 207–248. World Scientific (1998)
142. Tuia, D., Flamary, R., Courty, N.: Multiclass feature learning for hyperspectral image classification: sparse and hierarchical solutions. ISPRS J. Photogramm. Remote Sens. **105**, 272–285 (2015)
143. Tuia, D., Moser, G.: Foreword to the special issue on data fusion in remote sensing. IEEE Geosci. Remote Sens. Mag. **3**(3), 6–7 (2015)
144. Tuia, D., Pacifici, F., Kanevski, M., Emery, W.: Classification of very high spatial resolution imagery using mathematical morphology and support vector machines. IEEE Trans. Geosci. Remote Sens. **47**(11), 3866–3879 (2009)
145. Urbach, E.R., Roerdink, J.B.T.M., Wilkinson, M.H.F.: Connected shape-size pattern spectra for rotation and scale-invariant classification of gray-scale images. IEEE Trans. Pattern Anal. Mach. Intell. **29**(2), 272–285 (2007)
146. Urbach, E.R., Wilkinson, M.H.F.: Shape-only granulometries and grey-scale shape filters. In: Mathematical Morphology and Its Application to Signal and Image Processing - Proceedings of the 6th International Symposium on Mathematical Morphology, vol. 6, pp. 305–314 (2002)
147. Valero, S., Salembier, P., Chanussot, J.: Hyperspectral image representation and processing with binary partition trees. IEEE Trans. Image Process. **22**(4), 1430–1443 (2013)
148. Velasco-Forero, S., Angulo, J.: Classification of hyperspectral images by tensor modeling and additive morphological decomposition. Pattern Recogn. **46**(2), 566–577 (2013)
149. Voisin, A., Krylov, V., Moser, G., Serpico, S.B., Zerubia, J.: Supervised classification of multisensor and multiresolution remote sensing images with a hierarchical copula-based approach. IEEE Trans. Geosci. Remote Sens. **52**(6), 3346–3358 (2014)
150. Waske, B., Van Der Linden, S.: Classifying multilevel imagery from SAR and optical sensors by decision fusion. IEEE Trans. Geosci. Remote Sens. **46**(5), 1457–1466 (2008)
151. Wu, J., Jiang, Z., Luo, J., Zhang, H.: Composite kernels conditional random fields for remote-sensing image classification. Electron. Lett. **50**(22), 1589–1591 (2014)
152. Xia, J., Dalla Mura, M., Chanussot, J., Du, P., He, X.: Random subspace ensembles for hyperspectral image classification with extended morphological attribute profiles. IEEE Trans. Geosci. Remote Sens. **53**(9), 4768–4786 (2015)
153. Xia, J., Liao, W., Chanussot, J., Du, P., Song, G., Philips, W.: Improving random forest with ensemble of features and semisupervised feature extraction. IEEE Geosci. Remote Sens. Lett. **12**(7), 1471–1475 (2015)

154. Xu, Y., Carlinet, E., Géraud, T., Najman, L.: Efficient computation of attributes and saliency maps on tree-based image representations. In: Mathematical Morphology and Its Application to Signal and Image Processing - Proceedings of the 12th International Symposium on Mathematical Morphology, vol. 9082, pp. 693–704. Springer, Berlin (2015)
155. Xu, Y., Géraud, T., Najman, L.: Morphological filtering in shape spaces: applications using tree-based image representations. Proceedings of the 21st International Conference on Pattern Recognition **5**, 2–5 (2012)
156. Zhang, Y., Yang, H., Prasad, S., Pasolli, E., Jung, J., Crawford, M.: Ensemble multiple kernel active learning for classification of multisource remote sensing data. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. **8**(2), 845–858 (2015)
157. Zhang, Z., Pasolli, E., Crawford, M.M., Tilton, J.C.: An active learning framework for hyperspectral image classification using hierarchical segmentation. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. **9**(2), 640–654 (2016)
158. Zhao, W., Guo, Z., Yue, J., Zhang, X., Luo, L.: On combining multiscale deep learning features for the classification of hyperspectral remote sensing imagery. Int. J. Remote Sens. **36**(13), 3368–3379 (2015)
159. Zhong, Z., Fan, B., Duan, J., Wang, L., Ding, K., Xiang, S., Pan, C.: Discriminant tensor spectral-spatial feature extraction for hyperspectral image classification. IEEE Geosci. Remote Sens. Lett. **12**(5), 1028–1032 (2015)

Chapter 8
Change Detection in Multitemporal Images Through Single- and Multi-scale Approaches

Bruno Aiazzi, Francesca Bovolo, Lorenzo Bruzzone, Andrea Garzelli, Davide Pirrone and Claudia Zoppetti

Abstract This chapter presents an analysis of the current status and the challenges in change detection techniques for the analysis of multitemporal SAR images. Algorithms and methods based on validated statistical models for SAR data are investigated, which adopt advanced information-theoretic and multi-scale signal-processing methodologies. After a brief review of the recent literature on general change detection methods, the chapter investigates the specific problem of change detection in SAR images. The main properties of the change detection problem in SAR images are explored and discussed. Then, recent change detection techniques for high-resolution (HR) and very high-resolution (VHR) SAR data are presented and critically analyzed from the theoretical viewpoint. Finally, examples of application of these techniques to real problems are presented by using simulated image pairs and Enhanced Spotlight COSMO-SkyMed images.

B. Aiazzi
Institute of Applied Physics "Nello Carrara", National Research Council,
Via Madonna del Piano 10, 50019 Sesto Fiorentino, Firenze, Italy
e-mail: b.aiazzi@ifac.cnr.it

F. Bovolo · D. Pirrone
Fondazione Bruno Kessler,via Sommarive 18, 38123 Povo, Trento, Italy
e-mail: bovolo@fbk.eu

D. Pirrone
e-mail: pirrone@fbk.eu

L. Bruzzone
Department of Information Engineering and Computer Science, University of Trento,
via Sommarive 9, 38123 Povo, Trento, Italy
e-mail: lorenzo.bruzzone@unitn.it

A. Garzelli (✉) · C. Zoppetti
Department of Information Engineering and Mathematics, University of Siena,
via Roma 56, 53100 Siena, Italy
e-mail: andrea.garzelli@unisi.it

C. Zoppetti
e-mail: claudia.zoppetti@unisi.it

G. Moser and J. Zerubia (eds.), *Mathematical Models for Remote Sensing Image Processing*, Signals and Communication Technology,
https://doi.org/10.1007/978-3-319-66330-2_8

8.1 Introduction

Change detection (CD) in remote sensing is defined as the process of identifying changes in the features of the scene by means of the joint analysis of a pair of images acquired at different times over the same geographical area. CD has several applications in environmental monitoring, such as damage assessment or urban expansion monitoring.

Applications of change detection from spaceborne platforms progressively migrate from slowly changing phenomena (land-cover dynamic analysis, deforestation control) to rapidly mapping observations of natural or anthropic disasters, such as landslides, floods, earthquake damages, fires, and oil pollution. Polar orbiting satellites, however, do not provide adequate revisit time to monitor unpredictable and exceptional events. Therefore, making a direct comparison between post-event and pre-event data having identical acquisition parameters is almost unfeasible [29]. Hence, change detection is performed by comparing the first available data acquired after the event and previously archived acquisitions of the same scene.

One of the main challenges for a change detection algorithm is that the changes produced by the event under observation cannot be easily modeled. Actually, the same kind of event exhibits different signatures, depending on the region where it occurred, e.g., an urban or an agricultural area, and on the characteristics of the imaging sensor. Furthermore, when the time interval between two observations is large, changes to be identified are often mixed to seasonal or incidental changes that may be the majority, even if they usually have a minor extent and are often less relevant from the application viewpoint.

Several different approaches to change detection have been proposed in the literature [4, 10–22, 24–28, 30, 34–36, 43–46, 48–50, 53–55, 57–59, 62, 64–69].

Several examples can be found of (semi)supervised [18, 20, 22, 49, 64, 65, 69], and unsupervised [4, 10–17, 19, 21] methods as well. Labeled samples for each or some of the considered multitemporal acquisitions are required when supervised or semi/partially supervised methods are considered, whereas they are not for unsupervised ones. Thus the possibility of gathering reference samples for the training phase is an element that drives the kind of method to employ. Since the training sample collection is complex or even unfeasible, unsupervised approaches are often preferred. On the other side, application requirements should be considered as well. In fact, unsupervised methods do not provide a "from-to" information about the kind of change. Furthermore, unsupervised methods are specifically designed to handle multitemporal images acquired from either active SAR sensors [4, 11, 14, 25–27, 30, 34, 36, 43, 46, 48, 55, 57, 62], or optical passive ones [10, 12, 13, 16, 17, 19, 21, 35, 44, 50, 53, 54, 58, 59].

Other methods are more general and are able to handle multi-sensor information [15, 18, 20, 22, 49, 50, 64–69]. However, due to the scarce sensitivity of SAR to atmospheric and weather conditions, the available post-event data are likely to be SAR images [37]. Furthermore, the potentials of SAR sensors in change detection applications are strengthened by the high spatial resolution and the short revisit time

provided by the new generation SAR-based missions, such as COSMO-SkyMed (CSK), TerraSAR-X (TSX), Radarsat Constellation Mission (RCM), and Sentinel-1. The improvement in spatial resolution, which can reach 1 m for Spotlight products, is of fundamental importance in case of urban or suburban scenes [56]. In addition, the four-satellite constellation of the CSK system increases the possibility of monitoring the temporal evolution of an environmental disaster effectively. A worst-case minimum revisit time of 12 h is guaranteed.

This chapter starts with an analysis of the state of the art considering change detection methods for both optical passive and SAR active images. After that, attention is devoted to recent change detection techniques for high-resolution (HR) and very high resolution (VHR) SAR, with specific attention to the trade-off between effective speckle reduction from SAR data (see also Chaps. 4 and 5) and good preservation of the fine spatial details provided by the new generation of spaceborne SAR missions. Both simulated data and COSMO-SkyMed image pairs are considered for experimental evaluation and performance comparison among single-scale and multi-scale approaches.

8.2 State of the Art

8.2.1 Change Detection in Multitemporal Spaceborne Images

As mentioned in the introduction, the literature is plenty of methods for change detection both for optical passive and active SAR images. At a given level of abstraction, most of them follows similar philosophies. However, they strongly differ in the implementation details. This is because the statistical model of the two kinds of data is different: Optical passive image processing relies on an additive Gaussian noise model, whereas SAR image processing relies on a multiplicative speckle noise model.

Among supervised methods, three macro groups can be identified: post-classification comparison [65], supervised direct multidate classification [49, 65], and compound classification [20, 22, 23, 64, 69]. Post Classification Comparison (PCC) (also referred to as delta classification [65]) performs change detection by comparing the classification maps obtained by classifying two images independently. Multitemporal images are independently classified, thereby minimizing the problem of radiometric calibration, but ground truth is required for each of them. Although PCC has been used in several applications extensively, its performance strongly depends on the classification accuracy of the classifier applied to each single image. Supervised direct multidate classification (DMC) [49, 65] characterizes pixels by stacking the feature vectors related to the images acquired at the two different times. Each class transition is considered as a single class, thus the training pixels should represent the proportions of all the transitions in the whole area of interest accurately. This represents a serious drawback as, in real applications, it is difficult to

obtain training sets with such characteristic. A more realistic approach is compound classification (CC) [22], since it allows the temporal correlation between images to be considered in the change detection process. When ground truth is not available for each multitemporal acquisition, partially supervised classifiers can be used. They are able to update the classifier parameters estimated based on the ground truth available for one multitemporal image and match them to the statistical properties of multitemporal images for which the ground truth is not available. These methods, recently referred to as domain adaptation (DA) methods [18, 24, 31], have been investigated with novel interest because of the use of active learning (AL) [38–40]. All the aforementioned methods are based on classifiers like Maximum Likelihood classifier [42, 61], Neural Networks [23, 51], Fuzzy Classifiers [9, 39, 71], and Support Vector Machines [33, 70], which are either the most widely used or the most effective ones (the reader is referred to the literature for more details on the behaviors and mathematical details of each single classifier. An example of multitemporal classification of optical images is shown in Chap. 7). Because of this, such approaches are intrinsically suitable to process data acquired from either passive optical or active SAR systems, as well as to solve multi-sensor/multisource data problems. This becomes even more true when distribution-free non-parametric classifiers are considered.

Unfortunately, in several situations and applications, ground truth information cannot be collected, or the process becomes too expensive. In such situations, unsupervised methods become the only opportunity. This is the reason why the scientific community is still very active on this topic even if the literature is extensive. Once multitemporal images have been radiometrically and geometrically corrected, unsupervised change detection information extraction requires mainly two steps: (i) image comparison that results in a change feature (CF): this step aims at highlighting the presence of changes and accounts for the temporal correlation among acquisitions; (ii) analysis of the change feature. This step aims at isolating the change from the no-change information. The first step is the one that mostly depends on the kind of considered data.

When dealing with optical passive sensor images, comparison mainly relies on the difference operator. This is because the noise model in optical images is additive and the natural classes tend to have a Gaussian distribution. The simplest way to use the difference operator is to apply it to one or multiple corresponding spectral bands from multitemporal images [65], leading to the definition of Spectral Change Vectors (SCVs). The latter option is referred to as Change Vector Analysis and has been effectively employed with multispectral and hyperspectral images, and low to high-resolution images as well [12, 16, 19, 53, 54]. Under the assumption of Gaussian-distributed natural classes and being the difference a linear operator, classes of change and no change in the SCV feature space result to be Gaussian distributed as well [12]. Non-linear features are commonly extracted from SCVs [12, 53, 54, 72] like the magnitude and direction variables. The magnitude of changed samples presents significantly higher values than those of pixels associated with unchanged areas [19, 65]. Thus, the magnitude allows for a simple binary detection separating change and no change. On the other side, the direction variable is highly relevant for distinguishing among different kinds of changes as they assume preferred directions

[12, 53, 54, 65]. The difference operator can be applied in feature spaces other than the original spectral band one. Examples can be found that apply it to posterior probabilities [28], vegetation indexes [65], Tasselled Cap Transformation features [66], Multivariate Alteration Detection features [59], non-linear combinations of spectral bands, etc.

8.2.2 SAR Change Detection

When dealing with SAR images the commonly accepted noise model is multiplicative. Under this assumption, it is possible to show that after subtraction the statistical distribution of the resulting image depends on both the relative change between the intensity values in the two images and a reference intensity value (i.e., the intensity before or after the change). This leads to a higher change detection error for changes occurred in high-intensity regions of the image compared to that in low-intensity regions. Thus, the ratio operator (Image Rationing) [65] is more indicated for SAR multitemporal image comparison since its distribution depends only on the relative change in the average intensity between the two dates and not on a reference intensity level [8, 60]. Furthermore, it allows to reduce common multiplicative-error components [60]. In the literature, the ratio image is usually expressed in a logarithmic scale. Thus the log-ratio operator is typically preferred [8, 30, 36, 60, 62]. Another set of comparison operators widely used with SAR (but valid for optical data as well [57]) is the one based on the use of information theoretical similarity measures: the Kullback–Leibler (KL) divergence [48], the Mutual Information [4], and combinations of them. Recently the multi-scale/-resolution concept has been introduced in the multitemporal image analysis. This need emerged because of the complexity of SAR data and because of the intrinsic multi-resolution information available in the images acquired by the new generation high-resolution sensors. To properly model multi-scale/-resolution information, different approaches have been used. Among the others, we recall the Wavelet decomposition [11, 27], the Contourlet transform [52], and multi-scale feature profiles computed on varying windows size (Sect. 8.2.3.3), multi-scale segments [10, 45], and morphological profiles [35, 44]. More sophisticated approaches for the representation of multi-resolution information have been developed when very high spatial resolution (VHR) images are analyzed. They model the high-level semantic information in VHR images [14, 16, 50] and thus become intrinsically suitable for multi-sensor analysis [15].

The typical methodological approach for SAR images considers a direct pixel-based comparison of the two images, that generates a change feature (CF), which is taken as input for the decision step. Typically, many approaches in the literature consider an unsupervised thresholding of CF [4, 27]. Nevertheless, the analysis is affected by the multiplicative speckle noise present in the SAR images, and its compensation implies a degradation in terms of spatial resolution. In order to deal with this issue, CD methods were designed that achieve different trade-offs in terms

of both accuracy by the compensation of the speckle effect and preservation of the high-resolution geometrical information.

Scale-driven analysis [11, 48], among them, considers different scale levels in the CD analysis. It is based on the multi-scale decomposition of the CI image, on the selection of the reliable scales for each pixel and subsequent image fusion and decision. In particular, the multi-scale decomposition considers the use of two-dimensional filtering (e.g., stationary wavelet transform) on the image which applies, in both row-wise and column-wise, either low-pass or high-pass filtering.

8.2.3 Change Detection Methods for VHR SAR Images

Concerning VHR remote sensing images, the high geometrical information content requires both accurate definition and modeling of the concept of change, which is often associated with the specific goal of the application. The complexity is increased by the need to take into account all the specific issues related to the properties of VHR data. Standard unsupervised change detection techniques in the remote sensing literature often do not perform a detailed analysis of the concept of change. Usually, they compare two images acquired on the same geographical area at different times by assuming that their radiometric properties are similar except for the presence of changes occurred on the ground [16].

When application-oriented prior information is not available, change can be detected from the image radiometric properties only. For VHR amplitude SAR images, a single-scale approach is rarely effective, while multi-scale approaches can improve the detection performance through the analysis of different scales of representation of the change signal, where each scale is characterized by a different trade-off between speckle reduction and preservation of geometrical details [11].

Let us consider two SAR images X_k, of size $I \times J$, acquired over the same geographical area at two different times t_k, with $k = 1, 2$. Let us assume that the bi-temporal images are co-registered, geo-referenced, and radiometrically corrected. Let ω_{nc}, ω_c be the set of classes associated with unchanged and changed pixels, respectively.

Here, we investigate the capabilities of both single-scale and multi-scale approaches in detecting changes in bi-temporal SAR acquisitions X_1 and X_2. Three selected algorithms are described in the following subsections.

8.2.3.1 Information-Theoretic Feature

The mean-shift information-theoretic change detection (MS-ITCD) method relies on a feature capturing the structural change between X_1 and X_2. It is robust to the statistical change that may be originated by speckle and co-registration inaccuracies. The method starts from the scatter plot of the amplitude levels in the two images and

applies the mean-shift (MS) algorithm to find the modes of the underlying bivariate distribution [4].

The rationale of the algorithm is that the negative of the logarithm of the probability of a mean amplitude level in one image conditional to the mean amplitude level of the same pixel in the other image measures the amount of information associated to the pixel change and hence the amount of change, which may be related to the conditional information of couples of symbols emitted by two information sources [32].

Let $x_1(i, j)$ and $x_2(i, j)$ be the symbols emitted by the two information sources X_1 and X_2, respectively, where $i = 0, \ldots I - 1$ and $j = 0, \ldots J - 1$. The average information content of the two sources is given by their entropy, $H(X_1)$ and $H(X_2)$. In general, a part of such information is common to the two sources. This common information is called *mutual information* and is a measure of the statistical dependency between X_1 and X_2, i.e.,

$$I(X_1; X_2) = H(X_1) - H(X_1|X_2) \tag{8.1}$$

or, equivalently,

$$I(X_1; X_2) = I(X_2; X_1) = H(X_2) - H(X_2|X_1) \tag{8.2}$$

where $H(X_2|X_1)$ is the conditional entropy of X_2 to X_1 and represents the fraction of $H(X_2)$ that cannot be inferred from the knowledge of the reference source X_1, because it is due to unpredictable changes.

Given the conditional information between $x_2(m, n)$ and $x_1(m, n)$, that is,

$$I(x_2(i, j)|x_1(i, j)) \triangleq -\log[p(x_2|x_1)] \tag{8.3}$$

the conditional entropy is the expected value of (8.3),

$$H(X_2|X_1) \triangleq -\sum_{X_1}\sum_{X_2} p(x_1, x_2) \log[p(x_2|x_1)] \tag{8.4}$$

where $p(x_1, x_2)$ and $p(x_2|x_1)$ are the joint probabilities of $x_1(i, j)$ and $x_2(i, j)$ and the conditional probabilities of $x_2(i, j)$ to $x_1(i, j)$, respectively.

The method, which features a fast version of mean-shift (MS), and its earlier version are summarized in the following procedure:

1. Given two co-registered amplitude SAR images $x_1(i, j)$ and $x_2(i, j)$ taken on the same scene at different times, estimate their local means at each pixel, $\overline{x}_1(i, j)$ and $\overline{x}_2(i, j)$, over a $(2r + 1) \times (2r + 1)$ sliding window with Gaussian weighting.

2. Cross-calibrate $\bar{x}_2$ over $\bar{x}_1$ by matching the global mean and variance of the former to those of the latter; hereafter, let $\bar{x}_2$ denotes the histogram-matched version of $\bar{x}_2$.
3. Scale the values of both $\bar{x}_1$ and $\bar{x}_2$ by $\max_{i,j}\{\bar{x}_1, \bar{x}_2\}$; hereafter, let $\bar{x}_1$ and $\bar{x}_2$ denote the scaled version of the former $\bar{x}_1$ and $\bar{x}_2$.
4. Draw the scatter plot of $\overline{x}_2(i, j)$ against $\overline{x}_1(i, j)$; thus, the scatter plot is contained in a square of unity side and unchanged pixels lie along its main diagonal.
5. Multiply $\bar{x}_1$ and $\bar{x}_2$ by an integer L, which will denote the size of the 2D histogram obtained from the binning of the scatter plot. Hereafter, $(\bar{x}_1, \bar{x}_2) \in \{[0, L] \times [0, L]\}$.
6. Calculate the $L \times L$ joint histogram $h(m, n), m = 0, \ldots, L-1, n = 0, \ldots, L-1$, by counting all points $(\bar{x}_1, \bar{x}_2)$ such that $n < \bar{x}_1 \leq n+1$ and $m < \bar{x}_2 \leq m+1$.
7. Estimate the discrete joint probability density function (PDF), $p(m, n) = p(\lfloor \bar{x}_2 \rfloor, \lfloor \bar{x}_1 \rfloor)$, by normalizing $h(m, n)$ to the overall number of points and convolving it by a normalized triangular kernel of length $(2t + 1)$, according to Parzen window method.
8. Divide $p(m, n)$ by its maximum along $\bar{x}_1$, $\max_n p(m, n)$, so that its value is one, and hence the logarithm is zero (no change) when $p(m, n)$ attains its maximum over n:
$$q(m|n) = \frac{p(m, n)}{\max_n p(m, n)} = p(m|n) \cdot \frac{p(n)}{p_{\max_n}(m)}. \tag{8.5}$$
9. Pre-calculate a lookup table (LUT) of the information-theoretic change detection (ITCD) feature for each pair of (m, n) that indexes $q(m|n)$ as:
$$\mathcal{C}(m, n) = -\log\{q(m|n)\} \tag{8.6}$$
10. To calculate a map of plain ITCD feature, for each pixel (i, j), calculate $\bar{x}_1 = \bar{x}_1(i, j)$ and $\bar{x}_2 = \bar{x}_2(i, j)$, as in Step 5, then $\text{ITCD}(i, j) = \mathcal{C}(\lfloor \bar{x}_2 \rfloor, \lfloor \bar{x}_1 \rfloor)$.
11. To calculate a map of MS-enforced ITCD (MS-ITCD) feature, the MS clustering algorithm is applied to the "binned" scatter plot obtained at the end of Step 5, with a uniform kernel of radius R,
    - perform migration of the scatterpoints belonging to an original bin (m, n): start from the center of the bin and move all scatterpoints at the same time toward the center of the attracting cluster.
    - let $(\mathbf{m}, \mathbf{n})$ denote the integer valued stop coordinates of MS applied to the bin (m, n); the change feature $\mathcal{C}(\mathbf{m}, \mathbf{n})$ is associated with all scatterpoints originally belonging to (m, n).
    - for each pixel (i, j), calculate the bin (m, n) in which the pixel falls, replace (m, n) with $(\mathbf{m}, \mathbf{n})$ found through MS, set $\text{MS-ITCD}(i, j) = \mathcal{C}(\mathbf{m}, \mathbf{n})$.

The effect of MS is moving the scatterpoints contained in each bin toward the attracting center corresponding to a mode of the underlying PDF, as defined at Step 7. Unchanged pixels produce scatterpoints that are likely to be moved toward one of

the modes along with the main diagonal; conversely, changed pixels will be moved toward one of the modes far from the main diagonal.

The main difference of the MS-enforced ITCD, originally introduced in [3] from its earlier version, ITCD, [2, 5, 6], is that the presence of MS makes the feature to follow a *clustered* approach: The information-theoretic feature is calculated from the values of conditional probabilities roughly corresponding to the modes of the joint PDF.

The resulting MS-ITCD feature is considered and tested here in two configurations: as an example of high-performance single-scale change feature, and as the change feature adopted by the multi-scale strategy described in the next section.

8.2.3.2 Multi-scale CD Strategy Based on Wavelet Decomposition

A detailed description of the scale-driven CD approach, originally presented in [11], is provided in this section. The technique takes as input X_1 and X_2 and consists in four main steps: (i) change feature extraction by means of image comparison; (ii) multi-resolution decomposition; (iii) adaptive scale identification on the basis of local statistics of both the full and lower resolution data; and (iv) adaptive fusion based on the optimal scale level and generation of the final CD map. A general block scheme of the approach is represented in Fig. 8.1.

In the first step, image comparison is performed to compute a change feature (CF) that highlights backscattering variations. Among the operators presented in the literature, the log-ratio X_{LR} (see Eq. (8.7)) is used here to illustrate the method [8, 11, 63]. However the method can be applied to other CFs as well (e.g., ratio [7], KL divergence [48], difference). Accordingly, in the experiments, results will be illustrated both on the log-ratio and the ITCD features.

$$X_{LR} = \log \frac{X_2}{X_1} = \log X_2 - \log X_1. \tag{8.7}$$

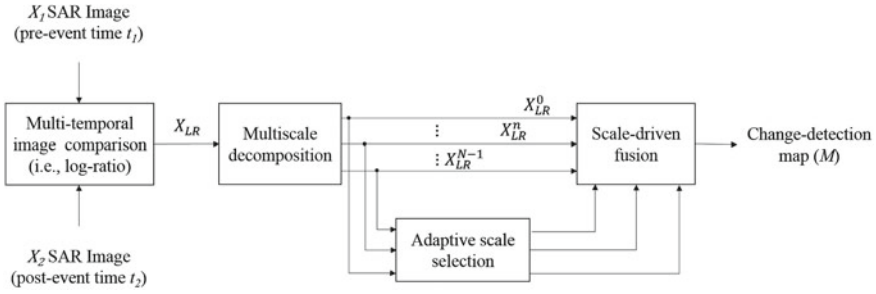


Fig. 8.1 Block scheme for the wavelet-based approach

The choice of the log-ratio allows to reduce the effect of the speckle noise and to have a statistical distribution of the CF centered on the zero value, with the two classes of interest assuming each a more symmetrical distribution.

The multitemporal information in X_{LR} is still affected by residual undesired speckle. Therefore, the second step aims at reducing the residual speckle effect on the log-ratio image, while preserving the geometrical details. For this reason, a decomposition of the log-ratio image at different scale levels is computed by creating a set of images $X_{ms} = \left\{X_{LR}^0, ..., X_{LR}^{N-1}\right\}$, where X_{LR}^n indicates the n-th level decomposition level, $n = 0, 1, \ldots, N-1$. To this end, a dyadic decomposition is applied, so that the scale corresponding to each resolution level is given by 2^n, and the image for $n = 0$ corresponds to the original log-ratio image. Among the possible approaches presented in the literature for the two-dimensional image decomposition, such as Laplacian pyramid decomposition [47] or recursively upsampled bicubic filter [41], we consider the two-dimensional stationary wavelet transform (2D-SWT) [11, 27].

This filtering approach applies level-dependent filters to the considered signal at each resolution level, by working separately along rows and columns, respectively. Typical filters for this kind of applications are 4th-order Daubechies filters [11].

This approach presents the advantage of avoiding down-sampling and possible aliasing impairments. At each step of the 2D-SWT, the image of the low-resolution component is taken as input and filtered, both row-wise and column-wise, with low-pass and high-pass filters, in order to separate lower resolution components (LL) and detail components on vertical (LH), horizontal (HL), and diagonal (HH) direction, respectively. They are defined as:

$$X_{LR}^{LL(n+1)}(i,j) = \sum_{p=0}^{D^n-1}\sum_{q=0}^{D^n-1} l^n[p]l^n[q]X_{LR}^{LL(n)}(i+p, j+q) \tag{8.8}$$

$$X_{LR}^{LH(n+1)}(i,j) = \sum_{p=0}^{D^n-1}\sum_{q=0}^{D^n-1} l^n[p]h^n[q]X_{LR}^{LL(n)}(i+p, j+q) \tag{8.9}$$

$$X_{LR}^{HL(n+1)}(i,j) = \sum_{p=0}^{D^n-1}\sum_{q=0}^{D^n-1} h^n[p]l^n[q]X_{LR}^{LL(n)}(i+p, j+q) \tag{8.10}$$

$$X_{LR}^{HH(n+1)}(i,j) = \sum_{p=0}^{D^n-1}\sum_{q=0}^{D^n-1} h^n[p]h^n[q]X_{LR}^{LL(n)}(i+p, j+q) \tag{8.11}$$

The filter coefficients at level $n + 1$ are obtained with a dilation of the coefficients of the filter at level n by a factor of 2. After the decomposition and by skipping the high resolution components, the approximation images at each scale level are retrieved by applying the inverse two-dimensional stationary wavelet (2D-ISWT) transform. Because of both the assumption on the additive noise model in the logarithmic scale and the computational cost of the processing, the wavelet strategy is applied directly on the log-ratio image, generating the set of images $X_{ms} = \left\{X_{LR}^0, \ldots, X_{LR}^{N-1}\right\}$.

Each of the wavelet outputs is used for generating a corresponding CD map, where the thresholds can be selected either manually or automatically, i.e., Bayesian approach based on the EM algorithm (see also Chaps. 4, 5, and 9), or Kittler-Illingworth thresholding, which is a computationally efficient solution to the problem of minimum error thresholding for normally distributed variables.

The final step of the algorithm is the fusion of the information in the products at different wavelet scales and the generation of the final CD map. Different fusion strategies are available in the literature, applying the fusion at the decision level and considering different choices for the thresholds. Fusion at the feature level on all reliable scales (FFL-ARS) [11] considers a reliable scale depending on the individual pixel and operates a fusion at *feature* level. Conversely, the fusion at the decision level on all reliable scales (FDL-ARS) [11] still considers a reliable scale depending on the individual pixel, but it operates the fusion at the *decision* level. The literature has proven that the best performance in terms of overall accuracy is obtained by the FFL-ARS approach, because of the best trade-off between the reduction of the speckle level and the details preservation. This fusion strategy is based on the generation of a new set of images, namely $\overline{X}_{ms} = \left\{\overline{X}_{ms}^{0}, ..., \overline{X}_{ms}^{N-1}\right\}$, derived from X_{ms}, in which each image $\overline{X}_{ms}^{n}$ is computed as an average of the wavelet decomposition up to the level n, as described in (8.12):

$$\overline{X}_{ms}^{n} = \frac{1}{n+1} \sum_{h=0}^{n} X_{LR}^{h}, \qquad n = 0, 1, \ldots, N-1 \tag{8.12}$$

In the FFL-ARS approach, for each pixel, reliable scale levels are determined according to whether the considered pixel belongs to either a border or a homogeneous region. In particular, for each of the scales it evaluates two coefficients: a global coefficient of variation (CV^n), defined on the whole image, and a local coefficient of variation ($LCV^n(i, j)$), defined on sliding window of user-defined size centered on the pixel (i, j).

These coefficients are expressed as:

$$LCV^n(i, j) = \frac{\sigma^n(i, j)}{\mu^n(i, j)} \tag{8.13}$$

$$CV^n = \frac{\sigma^n}{\mu^n} \tag{8.14}$$

where μ^n, σ^n represent the mean and the standard deviation on the corresponding areas, respectively.

The coefficient of variation cannot be computed on the multi-scale log-ratio images, so the computation of these two coefficients is done on the multi-scale ratio image sequence, derived by inverting the logarithm operation for each of the scale

levels. For a given pixel (i,j), the decomposition scale R_{ij} is defined as reliable if the following condition is satisfied for all the resolution levels ($l = 0, 1, ..., R_{ij}$):

$$M(i,j) \in \omega_k \Longleftrightarrow M^l(i,j) \in \omega_k, \quad \omega_k \in \{\omega_c, \omega_{nc}\}, \quad \forall l,\ 0 \le l \le R_{ij},\ R_{ij} \le N-1 \tag{8.15}$$

$$M^l(i,j) \in \omega_k \Longleftrightarrow LCV^l(i,j) \le CV^l, \quad \forall l,\ 0 \le l \le R_{ij}, \quad R_{ij} \le N-1 \tag{8.16}$$

The final CD map is obtained from the set of multi-resolution maps by applying a standard thresholding procedure to the fused images and recombining them by selecting the most reliable scale level for each pixel, as

$$M(i,j) = \begin{cases} \omega_{nc} & \text{if } x = \bar{X}_{ms}^{R_{ij}}(i,j) \le T^{R_{ij}} \\ \omega_c & \text{if } x = \bar{X}_{ms}^{R_{ij}}(i,j) > T^{R_{ij}} \end{cases} \tag{8.17}$$

where $T^{R_{ij}}$ is the decision threshold optimized for the considered fused image $\overline{X}_{ms}^{R_{ij}}$ and (i,j) is the spatial position of the considered pixel.

For the set $\overline{X}_{ms}$, the value at the reliable scale $\overline{X}_{ms}^{R_{ij}}(i,j)$ and the related threshold $T_{ms}^{R_{ij}}(i,j)$ are associated for each pixel (i,j). As described above, the threshold values, derived for the different wavelet levels, can be either manually or automatically set, according to any of the different strategies in the literature.

8.2.3.3 Combination of Multi-scale Change Features

Let r_w denote the bounded ratio image ($0 < r_w < 1$) computed from two co-registered amplitude SAR images X_1 and X_2 acquired on the same scene at different dates:

$$r_w = \min\left\{ \frac{\overline{X}_1^{(w)}}{\overline{X}_2^{(w)}}, \frac{\overline{X}_2^{(w)}}{\overline{X}_1^{(w)}} \right\}, \tag{8.18}$$

where $\overline{X}_k^{(w)}$ indicates the k-th image averaged over a $w \times w$ sliding window.

The bounded ratio image is computed for different odd window sizes in the interval $S = [w_{min}, w_{max}]$, and the resulting $N_w = (w_{max} - w_{min})/2 + 1$ multi-scale features are finally combined into the single geometric-mean bounded-ratio (GMBR) change feature R_S:

$$R_S = \left(\prod_{w \in S} r_w \right)^{1/N_w}. \tag{8.19}$$

The GMBR feature R_S has the important property of being intrinsically normalised, which is convenient for unsupervised clustering, it is easy to compute, robust to speckle impairments, and shows good capability of spatial detail preservation, thanks to its multi-scale nature.

The interval S of the window sizes should be selected according to the number of looks of the SAR images (or equivalently, their resolution), and possibly to the expected size of the regions of change. Typical values are $S = [5, 25]$ for 1-look data and $S = [3, 11]$ for 4-look data.

The GMBR feature can be properly clustered into the two classes of unchanged pixels, ω_{nc}, and changed pixels, ω_c, by applying the unsupervised K-means algorithm with $K = 2$. An example of R_S is reported in Fig. 8.4b, and the resulting change map obtained by K-means clustering is shown in Fig. 8.10b.

8.3 Experimental Results

This section presents the experimental results obtained on two different datasets, a simulated and a real one. In the next subsections, each dataset is described, and the performance analysis on four different change detection strategies is derived:

1. The MS-ITCD feature is first tested with optimal window size w and radius R (i.e., $w = 9$ pixels, $R = 30$ quantized amplitude levels) for best change mapping performance (see steps 1 and 11 of the MS-ITCD algorithm) by means of two-class k-means clustering.
2. The GMBR feature is computed and then clustered for change mapping.
3. The wavelet-based approach adopting the FFL-ARS fusion strategy driven by the log-ratio change feature, referred as FFL-ARS1 (see Fig. 8.1), is tested.
4. The FFL-ARS fusion strategy driven by the MS-ITCD feature with smaller window and radius ($w = 5$ pixels, $R = 10$ quantized amplitude levels), referred as FFL-ARS2, is finally applied.

8.3.1 Simulated Data

We have considered two different datasets of SAR images, on which different change events were simulated. In particular, these changes are related to four regions of backscattering increase and one of backscattering decrease. Two synthetic image

pairs with known patches with different shapes, sizes, and change levels have been produced from an optical remote sensing image, a panchromatic Ikonos image of Toulouse, France, with 12-bit dynamic range. Nakagami-distributed speckle patterns [1] have been generated with different equivalent number of looks L and spatial correlation values ρ_s, specifically, $L = 1$ and $\rho_s = 0.3$ for the first dataset, and $L = 4$ and $\rho_s = 0$ for the second dataset. It should be recalled that Nakagami-distributed speckle in the SAR amplitude domain is equivalent to Gamma-distributed speckle in the SAR intensity domain (see also Chaps. 4 and 5). The first dataset has 1m spatial resolution and 1m pixel spacing, thus simulating a CSK Enhanced Spotlight image pair, while the second dataset simulates Sentinel-1 Stripmap Mode images having 9 m spatial resolution and 4m pixel spacing.

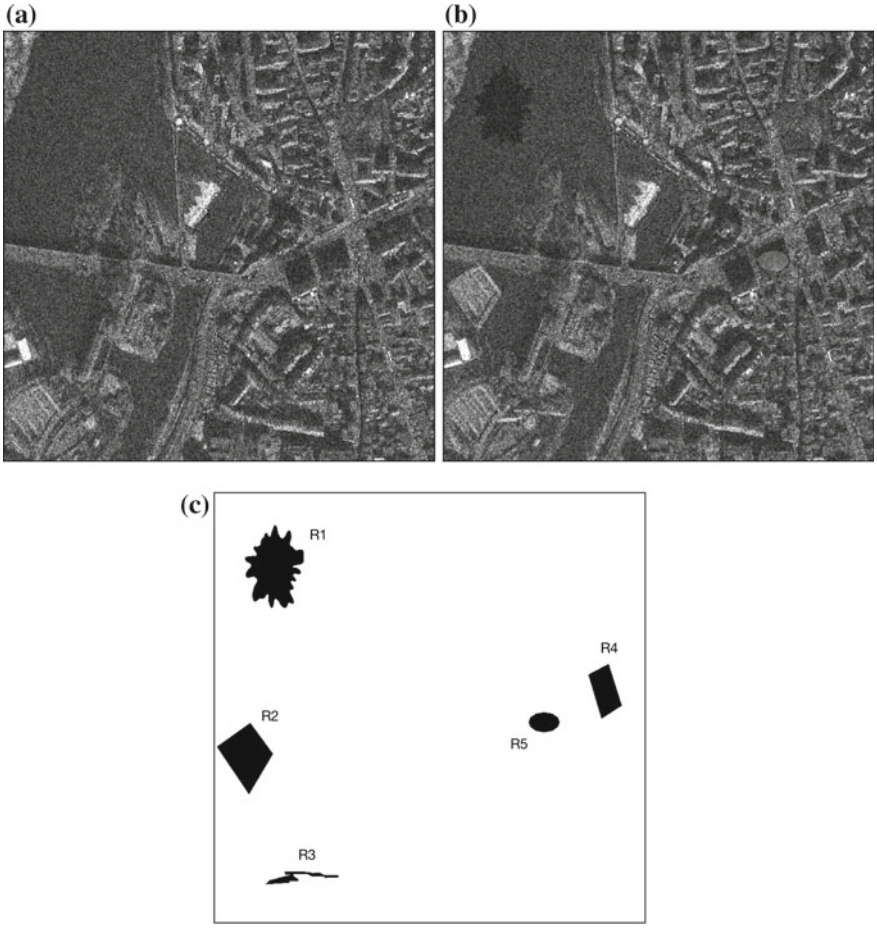


Fig. 8.2 **a**, **b**: Simulated 720 × 720 1-look image pair; ground truth change image (**c**)

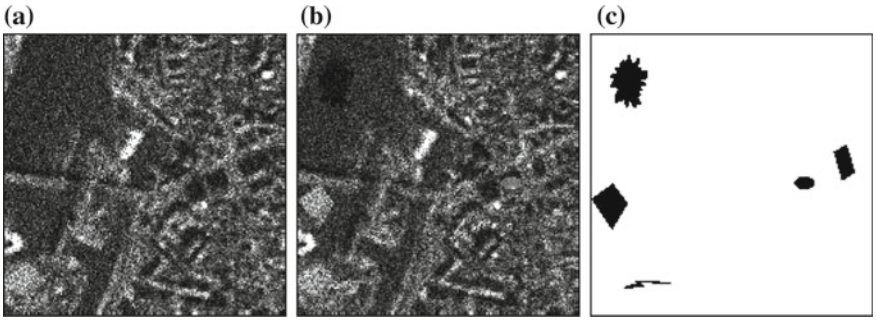


Fig. 8.3 **a**, **b**: Simulated 180 × 180 4-look image pair; ground truth change image (**c**)

The two datasets represent the same geographical area of about 0.5 km^2, through a 720 × 720 image pair for the 1-look dataset (Fig. 8.2), and a 180 × 180 image pair for the 4-look dataset (Fig. 8.3). The simulated change patches are regions with modified backscattering, specifically a 30% reduction of the amplitude level in the second date with respect to the first date in the R1 region, deterministic cover changes through pasting image values in the R2, R3, and R4 regions, and a constant increase of 80 amplitude levels on the R5 region (see Fig. 8.2c; corresponding regions are represented in Fig. 8.3c).

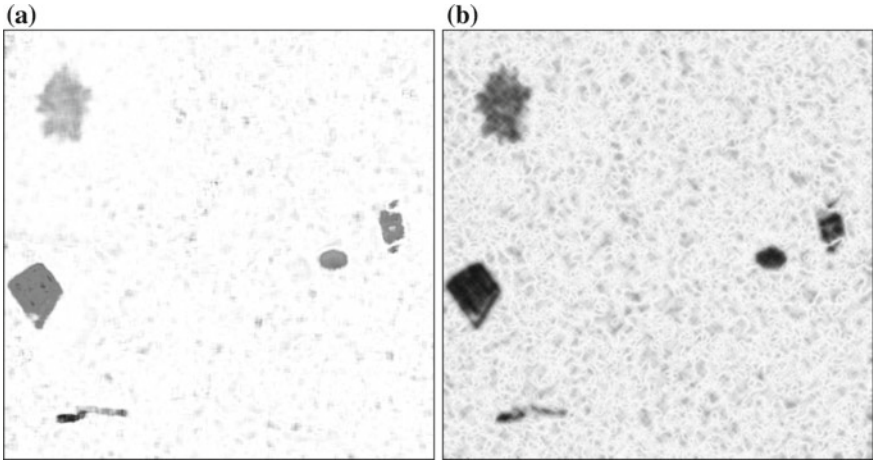


Fig. 8.4 Change features for the 1-look image pair: MS-ITCD (**a**); GMBR (**b**)

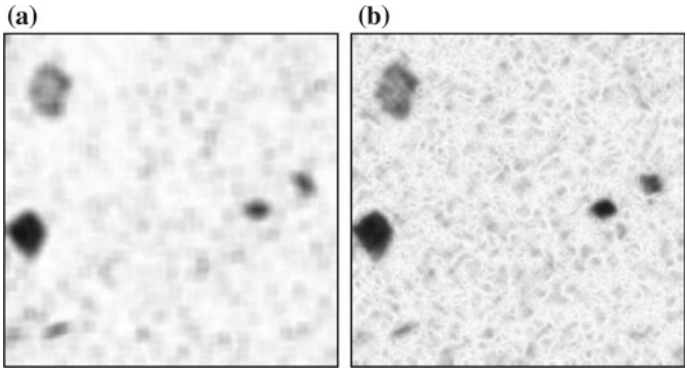


Fig. 8.5 Change features for the 4-look image pair: MS-ITCD (**a**); GMBR (**b**)

8.3.1.1 Quantitative Performance Assessment

The four change detection strategies are first compared in terms of the receiver operator characteristics (ROC). MS-ITCD and GMBR are directly compared in order to give evidence to the different characteristics of single-scale MS-ITCD and multi-scale GMBR features (Figs. 8.4 and 8.5).

FFL-ARS1 and FFL-ARS2 are not reported in the same graph since they cannot be considered as CD features, but as fusion strategies of multiple wavelet features. As previously stated, the FFL-ARS1 strategy relies on the log-ratio computation, while FFL-ARS2 is based on the MS-ITCD change index.

On the other hand, all the four CD methods can be straightly compared in terms of final binary CD maps, as reported in Sect. 8.3.1.2.

Figure 8.6 shows that MS-ITCD outperforms GMBR only for high values of false positive rate (FPR). However, since the optimal change maps, corresponding to the highest values of the Cohen's kappa coefficient (8.20), are obtained for lower values of FPR, as evidenced by Fig. 8.7, the best mapping performance for both 1-look and 4-look image pairs are provided by GMBR.

By comparing Fig. 8.6a, b, the GMBR ROC curve for the 1-look case is quite surprisingly higher than the ROC curve for the 4-look case (red curves in the two figures). This is due to the characteristics of GMBR which is more sensitive to the degradation of the spatial resolution rather than to the signal-to-noise reduction due to speckle.

For FFL-ARS1 and FFL-ARS2 experiments, ROC curves obtained at different wavelet decomposition levels are reported in Figs. 8.8 and 8.9, respectively. In order to have a suitable number of pixels at all decomposition levels, maximum decomposition levels $N = 5$ and $N = 3$ have been considered for the analysis of the 1-look and the 4-look image pairs, respectively.

For the FFL-ARS1 strategy, the set of images $\overline{X}_{ms}$, i.e., starting from the log-ratio image, has been derived. Each image in $\overline{X}_{ms}$ has been separately thresholded, and the

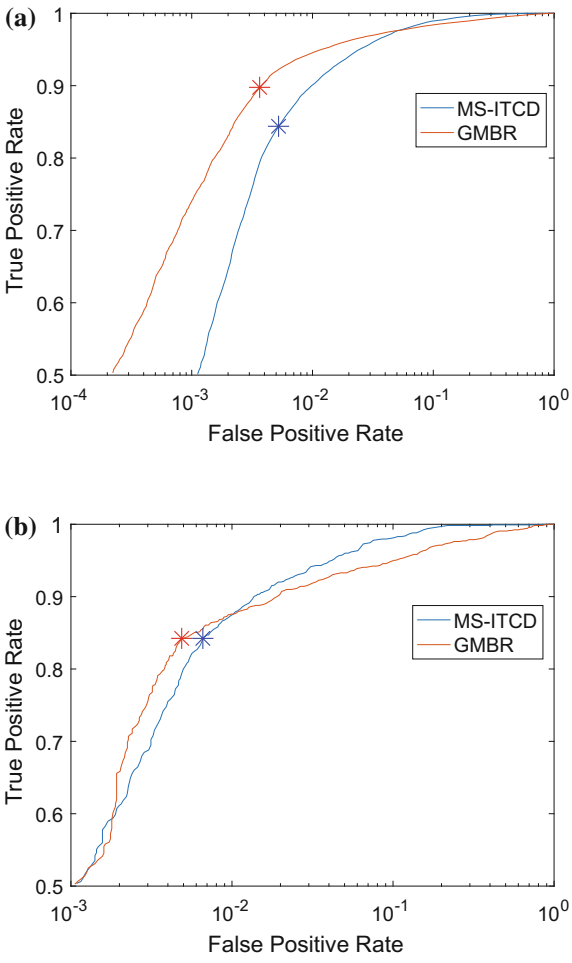


Fig. 8.6 ROC of the MS-ITCD and GMBR change features for the 1-look (**a**) and 4-look (**b**) simulated data. The asterisks highlight the optimal operating points (see also Fig. 8.7)

CD performance of each decomposition level has been evaluated by tracing ROCs. Figure 8.8 shows the ROC curves obtained at the different wavelet decomposition levels of the log-ratio feature. It should be noted that for the single-look case in Fig. 8.8a, as the wavelet decomposition level n increases, the ROC curves show higher true positive rate (TPR) and better CD performance, with an optimal number of decomposition levels equal to 4. This can be explained by the inclusion of the low-resolution information in the average products $\overline{X}_{ms}$. In these image components, the speckle effect is mitigated, and classification over homogeneous areas is improved.

At the highest decomposition level $N = 5$ this trend is not confirmed, due to an extreme degradation of the spatial resolution. Figure 8.8b shows that, for the 4-look case, by using only 2 or 3 levels of decomposition, depending on the required false positive rate, we get good CD performance. This is because of the lower geometric

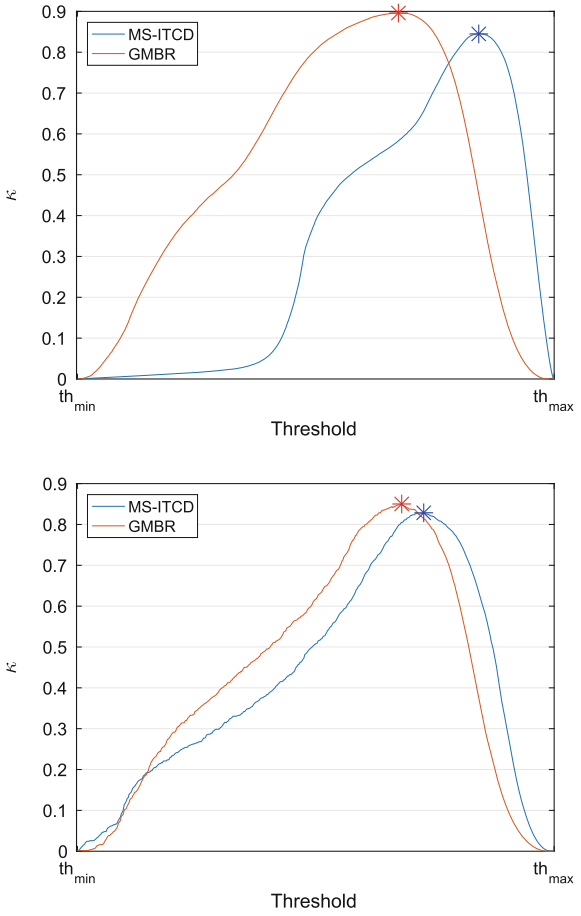


Fig. 8.7 Cohen's kappa of the feature-based change map with respect to the true change map as a function of the threshold applied to each feature. *Top* 1-look; *down* 4-look simulated data. The asterisks indicate the maximum values obtained by applying k-means clustering with $K = 2$ to GMBR (in *red*) and MS-ITCD (in *blue*)

resolution of the 4-look image with respect to the 1-look one. Thus, higher level of decomposition shows a too low geometrical resolution, with a degradation of the overall CD capabilities.

Concerning the FFL-ARS2 strategy, the set of images $\overline{X}_{ms}$, i.e., starting from the MS-ITCD feature, has been derived. Again, each $\overline{X}_{ms}$ image has been separately thresholded to produce a ROC curve. Figure 8.9 shows the ROC curves obtained at different wavelet decomposition levels by this method. For the single-look case in Fig. 8.9a, as the wavelet decomposition n increases, the ROC curves show higher TPR and better CD performance, with best performance for $N = 5$. Similarly to the FFL-ARS1 strategy, also for the FFL-ARS2 approach, Fig. 8.9b shows that, for the 4-look case, 2 or 3 levels of decomposition provide the best performance, depending on the required false positive rate. Therefore, the FFL-ARS algorithm can benefit from the different characteristics of the CD features at different scales.

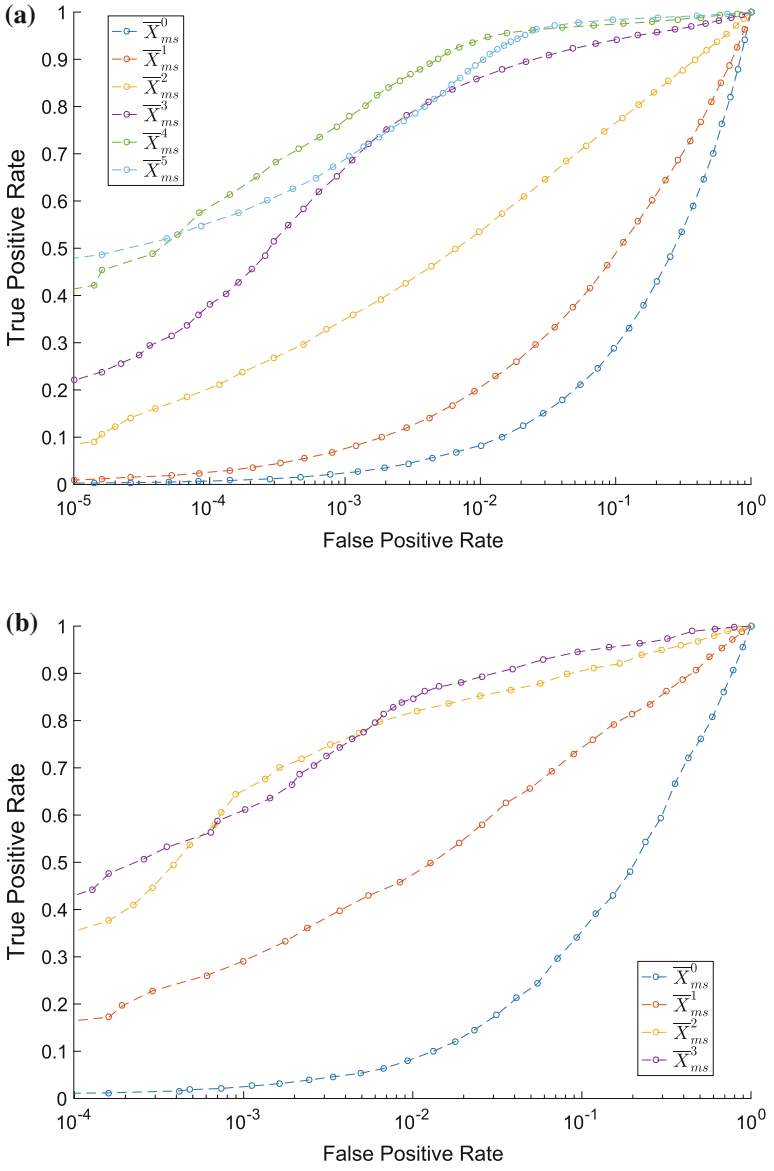


Fig. 8.8 ROC of the multi-scale change feature used by FFL-ARS1 for the 1-look (**a**) and 4-look (**b**) simulated data

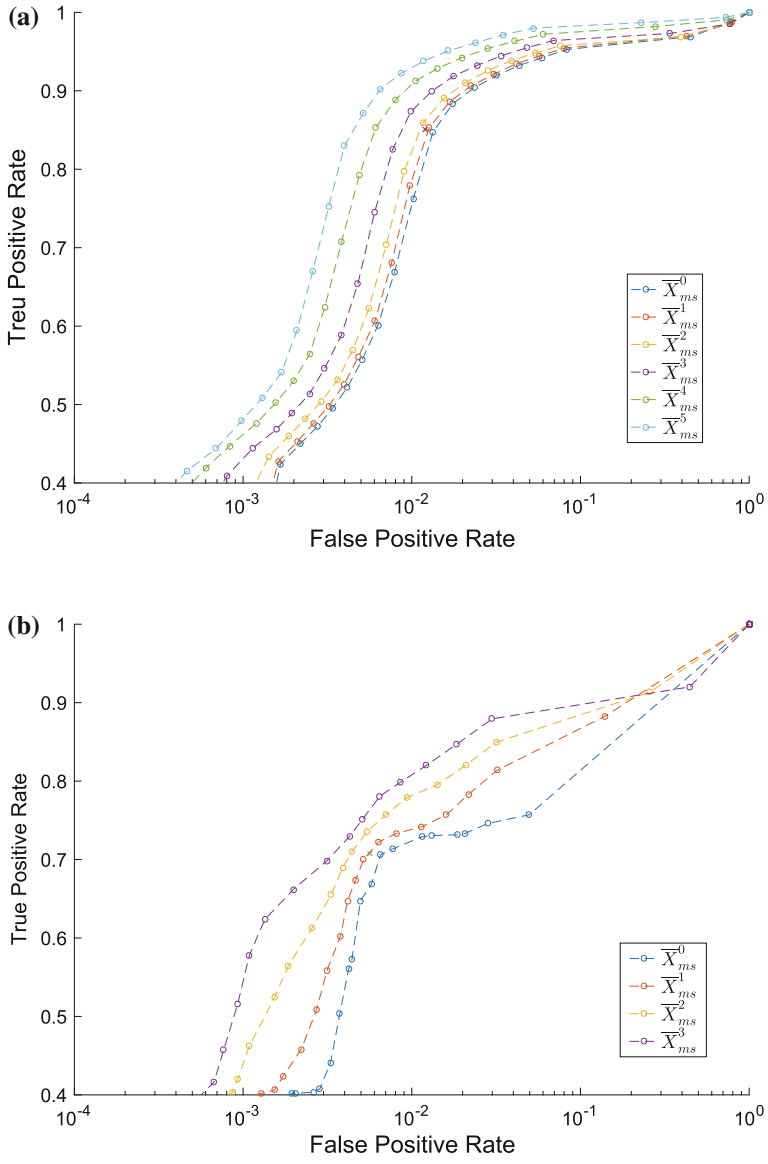


Fig. 8.9 ROC of the multi-scale change feature used by FFL-ARS2 for the 1-look (**a**) and 4-look (**b**) simulated data

Table 8.1 Cohen's kappa values of the CD maps obtained from the 1-look image pair

Algorithm	Cohen's kappa
MS-ITCD	0.860
GMBR	0.903
FFL-ARS1	0.612
FFL-ARS2	0.779

Since the ground truth is available for the two-simulated scenarios, the confusion matrix C can be computed to provide a quantitative performance assessment. The columns of the matrix represent the instances in the predicted classes (ω'_{nc}, i.e., no-change, in the first column, and ω'_c, *i.e.,* change, in the second column), while the rows represent the instances in the true classes (ω_{nc} or ω_c).

Starting from the confusion matrix, it is possible to compute the Cohen's kappa which compares the accuracy of the classification system to the accuracy of a random system:

$$\kappa = \frac{p_o - p_e}{1 - p_e} = 1 - \frac{1 - p_o}{1 - p_e} \tag{8.20}$$

where p_o is the overall accuracy and p_e is the accuracy of a random classifier. Differently from the confusion matrix, κ is a unique scalar value that provides a straightforward comparison among change maps obtained by different change detection algorithms.

For the 1-look image pair, we have:

$$\begin{aligned} C_{\text{MS-ITCD}} &= \begin{bmatrix} 497292 & 2337 \\ 2696 & 16075 \end{bmatrix} \quad C_{\text{GMBR}} = \begin{bmatrix} 498287 & 1342 \\ 2114 & 16657 \end{bmatrix} \\ C_{\text{FFL-ARS1}} &= \begin{bmatrix} 497672 & 1957 \\ 9412 & 9359 \end{bmatrix} \quad C_{\text{FFL-ARS2}} = \begin{bmatrix} 493331 & 6298 \\ 2478 & 16293 \end{bmatrix} \end{aligned} \tag{8.21}$$

corresponding to the Cohen's kappa values reported in Table 8.1.

We recall that the MS-ITCD result has been obtained with optimal window size and radius ($w = 9$ pixels, $R = 30$ quantised amplitude levels) for best change mapping performance by means of two-class k-means clustering, while the FFL-ARS2 fusion strategy on a 3×3 window has been driven by the MS-ITCD feature obtained with smaller window and radius, i.e., $w = 5$ pixels and $R = 10$ quantised amplitude levels.

For the 4-look image pair, we have:

$$\begin{aligned} C_{\text{MS-ITCD}} &= \begin{bmatrix} 31025 & 198 \\ 196 & 981 \end{bmatrix} \quad C_{\text{GMBR}} = \begin{bmatrix} 31097 & 126 \\ 223 & 954 \end{bmatrix} \\ C_{\text{FFL-ARS1}} &= \begin{bmatrix} 31154 & 69 \\ 510 & 667 \end{bmatrix} \quad C_{\text{FFL-ARS2}} = \begin{bmatrix} 31089 & 134 \\ 318 & 859 \end{bmatrix} \end{aligned} \tag{8.22}$$

corresponding to the Cohen's kappa values reported in Table 8.2.

The best performance, i.e., the highest κ values, are provided by the GMBR algorithm for both 1-look and 4-look data, thanks to an outstanding capability of rejecting false alarms with respect to FFL-ARS1 and, to a lesser extent, with respect to FFL-ARS2 and MS-ITCD. This advantage of GMBR is confirmed by Fig. 8.7,

Table 8.2 Cohen's kappa values of the CD maps obtained from the 4-look image pair

Algorithm	Cohen's kappa
MS-ITCD	0.826
GMBR	0.840
FFL-ARS1	0.689
FFL-ARS2	0.798

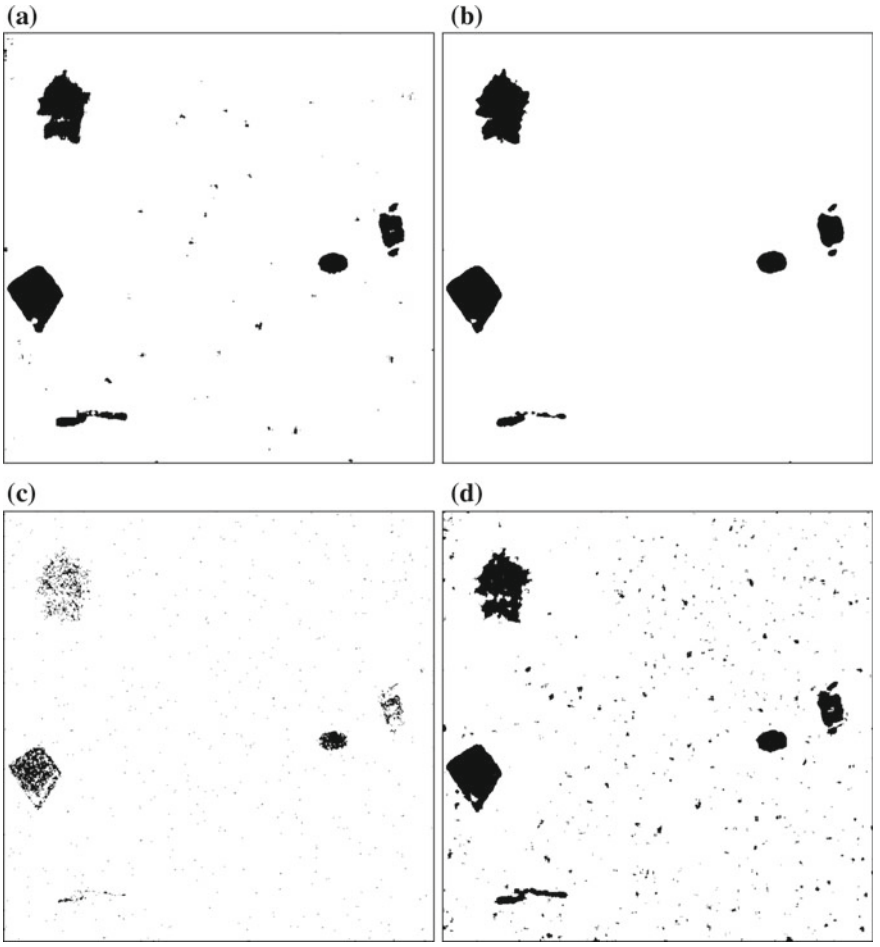


Fig. 8.10 Change maps for the 1-look case: MS-ITCD (**a**); GMBR (**b**); FFL-ARS1 (**c**); FFL-ARS2 (**d**)

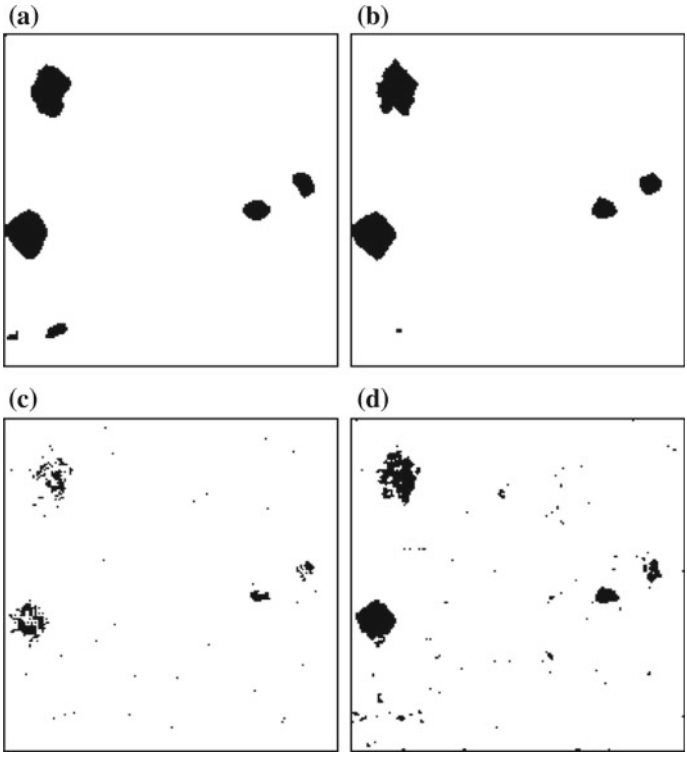


Fig. 8.11 Change maps from the 4-look image pair: MS-ITCD (**a**); GMBR (**b**); FFL-ARS1 (**c**); FFL-ARS2 (**d**)

which also shows that the K-means clustering provides the optimal threshold values corresponding to the maximum values of κ.

Concerning the FFL-ARS method, detection performance depends on the choice of the window size parameter for the CV computation. In particular, for the FFL-ARS1 case, the overall performance increases, both in terms of κ coefficient and misclassified pixels for large windows, while for FFL-ARS2 a small window size is preferred. In general, the FFL-ARS strategy aims at keeping the edge information of the changed regions. This is clear by observing the left central change in Fig. 8.10d when compared to the ground truth region R2 in Fig. 8.2c.

The advantage of adopting a robust change feature such as MS-ITCD instead of the log-ratio image in the FFL-ARS fusion strategy is evident for both 1-look and 4-look data, as shown by numerical and visual results.

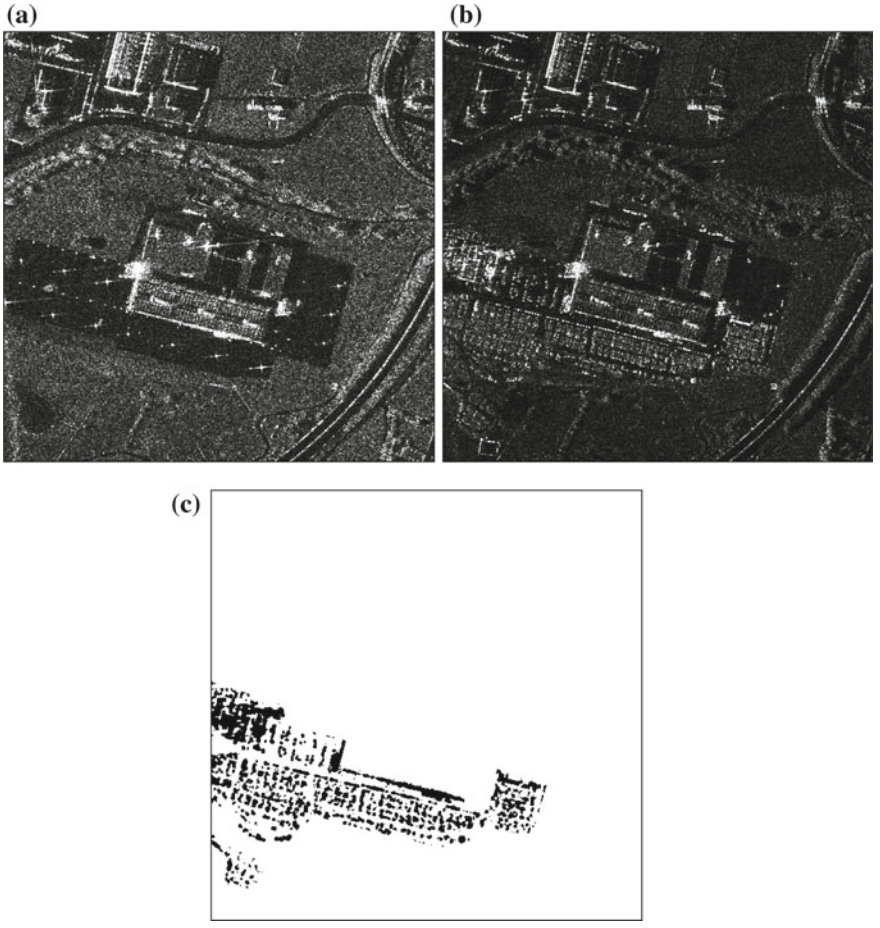


Fig. 8.12 Original pre- and post-event Spotlight acquisitions of L'Aquila test site: **a** April 5, 2009; **b**; September 12, 2009; manually generated ground truth of change, (**c**)

8.3.1.2 Qualitative Performance Assessment

Figure 8.10a and b show the final change maps computed by clustering, through K-means with K = 2, the single-scale feature MS-ITCD (Sect. 8.2.3.1) and the multi-scale feature GMBR (8.19), respectively. Both maps confirm the objective evaluation given by the confusion matrices and the κ values, with excellent detection capabilities in the 1-look case. The multi-scale nature of GMBR also provides a higher false alarm rejection, as shown in Fig. 8.10b.

The comparison between the change detection maps obtained with FFL-ARS1 and FFL-ARS2 (with $w = 5$ and $R = 10$) (Fig. 8.10c, d) points out that the former one presents several misclassification errors, while the latter is much more accurate. An

interesting characteristic of the FFL-ARS2 approach is its capability of preserving the spatial details of the changed regions, even better than the best performing GMBR algorithm.

Similar considerations apply for the change maps in the 4-look case reported in Fig. 8.11. The GMBR method provides the best change maps and shows very good false alarm rejection, while MS-ITCD seems to suffer from the poor spatial resolution of 4-look data. FFL-ARS, in its MS-ITCD driven version, and FFL-ARS2 can provide a good quality CD map, as in Fig. 8.11d, although at the expense of an increased false alarm rate.

8.3.2 COSMO-SkyMed Images

For the real dataset, two COSMO-SkyMed images have been considered and processed for assessing change detection capabilities in a true scenario. The considered change between the two acquisitions (April 5, 2009 in Fig. 8.12a and September 12, 2009 in Fig. 8.12b, i.e., before and after the destructive earthquake on April 9) is the construction of a tent camp set up for earthquake survivors near a shopping mall about 7 km west of the City Center. Both 1-look acquisitions have been taken with right look-side, ascending pass, HH polarization, and 58° incidence angle. The two images have 1m^2 pixel size and are 1000×1000 pixels.

A manually generated ground truth of the tent camp built after the earthquake is reported in Fig. 8.12c.

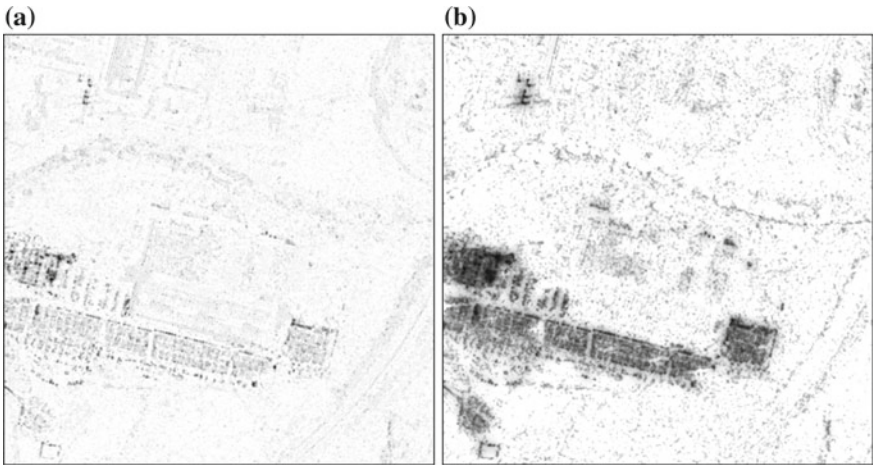


Fig. 8.13 Change features from the CSK image pair of Fig. 8.12: **a** MS-ITCD; **b** GMBR

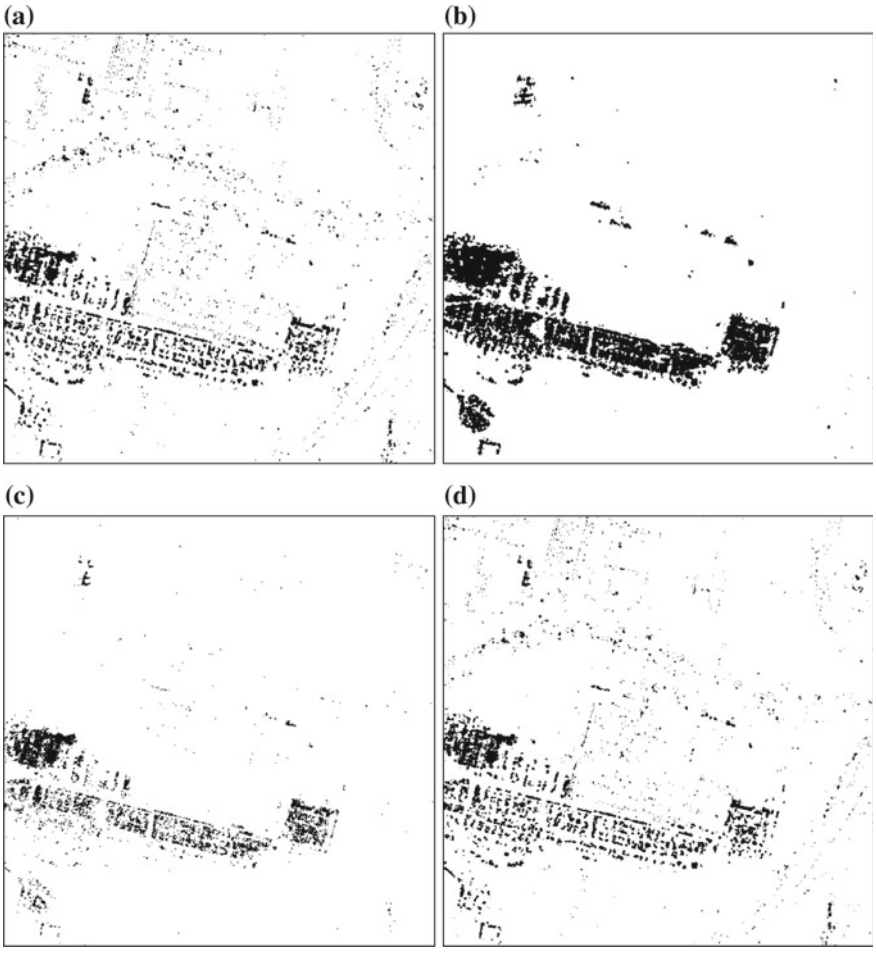


Fig. 8.14 Change maps obtained from the L'Aquila test site: **a** MS-ITCD; **b** GMBR; **c** FFL-ARS1; **d** FFL-ARS2

The MS-ITCD and GMBR change features computed on the image pair of Fig. 8.12 are reported in Fig. 8.13 showing similar responses to structural and statistical changes, but different dynamic ranges.

The final change maps obtained by applying MS-ITCD, GMBR, FFL-ARS1, and FFL-ARS2 are shown in Fig. 8.14. The two original images have been preprocessed to equalize their histograms, and the analysis has been focused on the regions with increased backscattering at the second date. All methods provide a clear description of the changed region with different results in terms of detection capability, false alarm rejection, and geometrical accuracy.

The analysis of the confusion matrices and the Cohen's kappa values provides an objective assessment of the characteristics of the three algorithms. The confusion

Table 8.3 Cohen's kappa values of the CD maps obtained from the CSK image pair

Algorithm	Cohen's kappa
MS-ITCD	0.685
GMBR	0.622
FFL-ARS1	0.527
FFL-ARS2	0.672

matrices are the following:

$$C_{\text{MS-ITCD}} = \begin{bmatrix} 934874 & 17202 \\ 12799 & 35125 \end{bmatrix} \quad C_{\text{FFL-ARS1}} = \begin{bmatrix} 940501 & 7172 \\ 30054 & 22273 \end{bmatrix}$$
$$C_{\text{FFL-ARS2}} = \begin{bmatrix} 936092 & 11581 \\ 18827 & 33500 \end{bmatrix} \quad C_{\text{GMBR}} = \begin{bmatrix} 922963 & 24710 \\ 15780 & 36547 \end{bmatrix} \tag{8.23}$$

which can be synthesized by the unique index κ in Table 8.3.

The relatively low κ values are due to inaccuracies of the ground truth which reports changes in the tent camp area only.

FFL-ARS2 and MS-ITCD show the highest κ values, thanks to the MS-ITCD accurate detection of distributed scatterers, which provides a very good preservation of small spatial features in the changed regions. FFL-ARS1 suffers from severe miss-detection due to its underlying log-ratio feature, while GMBR, although its change map appears clean and accurate at a first sight, is not capable of precisely locate the small distributed scatterers which characterize the change regions of the tested CSK image pair.

8.4 Concluding Remarks

The capabilities of both single-scale and multi-scale approaches of detecting changes from two-date SAR acquisitions have been investigated and experimentally assessed on 1-look and 4-look simulated images and on COSMO-SkyMed Spotlight SAR data. It has been shown that when application-oriented prior information is not available for modeling different kinds of changes, multi-scale approaches can be profitably applied to detect changes directly from the image radiometric properties at different dates. Among the tested algorithms, the single-scale MS-ITCD method has shown very accurate detection of distributed scatterers. The multi-scale FFL-ARS algorithm, when driven by advanced change features such as MS-ITCD, has evidenced good shape preservation, while the multi-scale GMBR algorithm has demonstrated the best trade-off, for both simulated and true data, between speckle reduction and preservation of geometrical details.

References

1. Achim, A., Kuruoglu, E.E., Zerubia, J.: SAR image filtering based on the heavy-tailed Rayleigh model. IEEE Trans. Image Process. **15**(9), 2686–2693 (2006)
2. Aiazzi, B., Alparone, L., Baronti, S., Garzelli, A., Nencini, F: Information-theoretic multitemporal features for change analysis from SAR images. In: Image and Signal Processing for Remote Sensing XIV, 7109, p. 71090S (2008)
3. Aiazzi, B., Alparone, L., Baronti, S., Garzelli, A., Zoppetti, C.: A robust change detection feature for COSMO-SkyMed detected SAR images. In: Proceedings of MultiTemp 2011, International Workshop on the Analysis of Multi-temporal Remote Sensing Images, pp. 125–128 (2011)
4. Aiazzi, B., Alparone, L., Baronti, S., Garzelli, A., Zoppetti, C.: Nonparametric change detection in multitemporal SAR images based on mean-shift clustering. IEEE Trans. Geosci. Remote Sens. **51**(4), 2022–2031 (2013)
5. Alparone, L., Aiazzi, B., Baronti, S., Garzelli, A.D: An information-theoretic feature for multitemporal analysis of SAR images. In: Proceedings of ESA-EUSC 2006: Image Information Mining for Security and Intelligence, ESA Workshop Proceedings Publication WPP-274, pp. 67–76 (2006)
6. Alparone, L., Aiazzi, B., Baronti, S., Garzelli, A., Nencini, F.: Robust change analysis of SAR data through information-theoretic multi temporal features. In: Proceedings of IEEE International Geoscience and Remote Sensing Symposium, pp. 3883–3886 (2007)
7. Ban, Y., Yousif, O.A.: Multitemporal spaceborne SAR data for urban change detection in China. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. **5**(4), 1087–1094 (2012)
8. Bazi, Y., Bruzzone, L., Melgani, F.: An unsupervised approach based on the generalized Gaussian model to automatic change detection in multitemporal SAR images. IEEE Trans. Geosci Remote Sens. **43**(4), 874–887 (2005)
9. Bezdek, J.C., Ehrlich, R., Full, W.: FCM: the fuzzy c-means clustering algorithm. Comput. Geosci. **10**(2–3), 191–203 (1984)
10. Bovolo, F.: A multilevel parcel-based approach to change detection in very high resolution multitemporal images. IEEE Geosci. Remote Sens. Lett. **6**(1), 33–37 (2009)
11. Bovolo, F., Bruzzone, L.: A detail-preserving scale-driven approach to change detection in multitemporal SAR images. IEEE Trans. Geosci. Remote Sens. **43**(12), 2963–2972 (2005)
12. Bovolo, F., Bruzzone, L.: A theoretical framework for unsupervised change detection based on change vector analysis in the polar domain. IEEE Trans. Geosci. Remote Sens. **45**(1), 218–236 (2007)
13. Bovolo, F., Marchesi, S., Bruzzone, L.: A framework for automatic and unsupervised detection of multiple changes in multitemporal images. IEEE Trans. Geosci. Remote Sens. **50**(6), 2196–2212 (2012)
14. Bovolo, F., Marin, C., Bruzzone, L.: A hierarchical approach to change detection in very high resolution SAR images for surveillance applications. IEEE Trans. Geosci. Remote Sens. **51**(4), 2042–2054 (2013)
15. Brunner, D., Lemoine, G., Bruzzone, L.: Earthquake damage assessment of buildings using VHR optical and SAR imagery. IEEE Trans. Geosci. Remote Sens. **48**(5), 2403–2420 (2010)
16. Bruzzone, L., Bovolo, F.: A novel framework for the design of change-detection systems for very-high-resolution remote sensing images. Proc. IEEE **101**(3), 609–630 (2013)
17. Bruzzone, L., Cossu, R.: An adaptive approach to reducing registration noise effects in unsupervised change detection. IEEE Trans. Geosci. Remote Sens. **41**(11), 2455–2465 (2003)
18. Bruzzone, L., Fernández Prieto, D.: A technique for the selection of kernel-function parameters in RBF neural networks for classification of remote-sensing images. IEEE Trans. Geosci. Remote Sens. **37**(2), 1179–1184 (1999)
19. Bruzzone, L., Fernández Prieto, D.: Automatic analysis of the difference image for unsupervised change detection. IEEE Trans. Geosci. Remote Sens. **38**(3), 1171–1182 (2000)

20. Bruzzone, L., Fernández Prieto, D.: Unsupervised retraining of a maximum likelihood classifier for the analysis of multitemporal remote sensing images. IEEE Trans. Geosci. Remote Sens. **39**(2), 456–460 (2001)
21. Bruzzone, L., Fernández Prieto, D.: An adaptive semiparametric and context-based approach to unsupervised change detection in multitemporal remote-sensing images. IEEE Trans. Image Process. **11**(4), 452–466 (2002)
22. Bruzzone, L., Serpico, S.B.: An iterative technique for the detection of land-cover transitions in multitemporal remote-sensing images. IEEE Trans. Geosci. Remote Sens. **35**(4), 858–867 (1997)
23. Bruzzone, L., Fernández Prieto, D., Serpico, S.B.: A neural-statistical approach to multitemporal and multisource remote-sensing image classification. IEEE Trans. Geosci. Remote Sens. **37**(3), 1350–1359 (1999)
24. Bruzzone, L., Cossu, R., Vernazza, G.: Detection of land-cover transitions by combining multidate classifiers. Pattern Recognit. Lett. **25**(13), 1491–1500 (2004)
25. Carincotte, C., Derrode, S., Bourennane, S.: Unsupervised change detection on SAR images using fuzzy hidden Markov chains. IEEE Trans. Geosci. Remote Sens. **44**(2), 432–441 (2006)
26. Celik, T., Ma, K.K.: Unsupervised change detection for satellite images using dual-tree complex wavelet transform. IEEE Trans. Geosci. Remote Sens. **48**(3), 1199–1210 (2010)
27. Celik, T., Ma, K.K.: Multitemporal image change detection using undecimated discrete wavelet transform and active contours. IEEE Trans. Geosci. Remote Sens. **49**(2), 706–716 (2011)
28. Chen, J., Chen, X., Cui, X., Chen, J.: Change vector analysis in posterior probability space: a new method for land cover change detection. IEEE Trans. Geosci. Remote Sens. **8**(2), 317–321 (2011)
29. Chini, M., Pulvirenti, L., Pierdicca, N.: Analysis and interpretation of the COSMO-SkyMed observations of the 2011 Japan tsunami. IEEE Geosci. Remote Sens. Lett. **9**(3), 467–471 (2012)
30. Cihlar, J., Pultz, T.J., Gray, A.: Change detection with synthetic aperture radar. Int. J. Remote Sens. **13**(3), 401–414 (1992)
31. Cossu, R., Chaudhuri, S., Bruzzone, L.: A spatial-contextual partially supervised classifier based on Markov random fields. IEEE Geosci. Remote Sens. Lett. 352–356 (2005)
32. Cover, T.M., Thomas, J.A.: Elements of Information Theory, 2nd edn. Wiley, New York (2006)
33. Cristianini, N., Shawe-Taylor, J.: An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods. Cambridge University Press, Cambridge (2000)
34. Cui, S., Datcu, M.: Statistical wavelet subband modeling for multi-temporal SAR change detection. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. **5**(4), 1095–1109 (2012)
35. Dalla Mura, M., Benediktsson, J.A., Bovolo, F., Bruzzone, L.: An unsupervised technique based on morphological filters for change detection in very high resolution images. IEEE Geosci. Remote. Sens. Lett. **5**(3), 433–437 (2008)
36. Dekker, R.J.: Speckle filtering in satellite SAR change detection imagery. Int. J. Remote Sens. **19**(6), 1133–1146 (1998)
37. Dekker, R.J.: High-resolution radar damage assessment after the earthquake in Haiti on 12 january 2010. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. **4**(4), 960–970 (2011)
38. Demir, B., Bovolo, F., Bruzzone, L.: Detection of land-cover transitions in multitemporal remote sensing images with active-learning-based compound classification. IEEE Trans. Geosci. Remote Sens. **50**(5), 1930–1941 (2012)
39. Demir, B., Bovolo, F., Bruzzone, L.: Classification of time series of multispectral images with limited training data. IEEE Trans. Image Process. **22**(8), 3219–3233 (2013)
40. Demir, B., Bovolo, F., Bruzzone, L.: Updating land-cover maps by classification of image time series: a novel change-detection-driven transfer learning approach. IEEE Trans. Geosci. Remote Sens. **51**(1), 300–312 (2013)
41. Dippel, S., Stahl, M., Wiemker, R., Blaffert, T.: Multiscale contrast enhancement for radiographies: Laplacian pyramid versus fast wavelet transform. IEEE Trans. Med. Imaging **21**(4), 343–353 (2002)
42. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification. Wiley, New York (2012)

43. Erten, E., Reigber, A., Ferro-Famil, L., Hellwich, O.: A new coherent similarity measure for temporal multichannel scene characterization. IEEE Trans. Geosci. Remote Sens. **50**(7), 2839–2851 (2012)
44. Falco, N., Dalla Mura, M., Bovolo, F., Benediktsson, J.A., Bruzzone, L.: Change detection in VHR images based on morphological attribute profiles. IEEE Geosci. Remote Sens. Lett. **10**(3), 636–640 (2013)
45. Garzelli, A., Zoppetti, C.: A segmentation-based approach to SAR change detection and mapping. In: Proceedings of SPIE 10004, Image and Signal Processing for Remote Sensing XXII, pp. 1000, 410–1000, 410–10 (2016)
46. Gueguen, L., Soille, P., Pesaresi, M.: Change detection based on information measure. IEEE Trans. Geosci. Remote Sens. **49**(11), 4503–4515 (2011)
47. Hay, G.J., Castilla, G., Wulder, M.A., Ruiz, J.R.: An automated object-based approach for the multiscale image segmentation of forest scenes. Int. J. Appl. Earth Obs. Geoinf. **7**(4), 339–359 (2005)
48. Inglada, J., Mercier, G.: A new statistical similarity measure for change detection in multitemporal SAR images and its extension to multiscale change analysis. IEEE Trans. Geosci. Remote Sens. **45**(5), 1432–1445 (2007)
49. Jeon, B., Landgrebe, D.A.: Classification with spatio-temporal interpixel class dependency contexts. IEEE Trans. Geosci. Remote Sens. **30**(4), 663–672 (1992)
50. Klaric, M.N., Claywell, B.C., Scott, G.J., Hudson, N.J., Sjahputera, O., Li, Y., Barratt, S.T., Keller, J.M., Davis, C.H.: GeoCDX: an automated change detection and exploitation system for high-resolution satellite imagery. IEEE Trans. Geosci. Remote Sens. **51**(4), 2067–2086 (2013)
51. Kosko, B.: Neural Networks and Fuzzy Systems: A Dynamical Systems Approach to Machine Intelligence. Prentice Hall, Upper Saddle River (1992)
52. Li, S., Fang, L., Yin, H.: Multitemporal image change detection using a detail-enhancing approach with nonsubsampled contourlet transform. IEEE Geosci. Remote Sens. Lett. **9**(5), 836–840 (2012)
53. Liu, S., Bruzzone, L., Bovolo, F., Du, P.: Hierarchical unsupervised change detection in multitemporal hyperspectral images. IEEE Trans. Geosci. Remote Sens. **53**(1), 244–260 (2015)
54. Liu, S., Bruzzone, L., Bovolo, F., Zanetti, M., Du, P.: Sequential spectral change vector analysis for iteratively discovering and detecting multiple changes in hyperspectral images. IEEE Trans. Geosci. Remote Sens. **53**(8), 4363–4378 (2015)
55. Marin, C., Bovolo, F., Bruzzone, L.: Building change detection in multitemporal very high resolution SAR images. IEEE Trans. Geosci. Remote Sens. **53**(5), 2664–2682 (2015)
56. Mason, D.C., Speck, R., Devereux, B., Schumann, G.J.P., Neal, J.C., Bates, P.D.: Flood detection in urban areas using TerraSAR-X. IEEE Trans. Geosci. Remote Sens. **48**(2), 882–894 (2010)
57. Mercier, G., Moser, G., Serpico, S.B.: Conditional copulas for change detection in heterogeneous remote sensing images. IEEE Trans. Geosci. Remote Sens. **46**(5), 1428–1441 (2008)
58. Moser, G., Angiati, E., Serpico, S.B.: Multiscale unsupervised change detection on optical images by Markov random fields and wavelets. IEEE Geosci. Remote Sens. Lett. **8**(4), 725–729 (2011)
59. Nielsen, A.A.: The regularized iteratively reweighted mad method for change detection in multi-and hyperspectral data. IEEE Trans. Image Process. **16**(2), 463–478 (2007)
60. Oliver, C., Quegan, S.: Understanding Synthetic Aperture Radar Images. SciTech Publishing (2004)
61. Richards, J.A.: Remote Sensing Digital Image Analysis, 5th edn. Springer, Berlin (2013)
62. Rignot, E.J., van Zyl, J.J.: Change detection techniques for ERS-1 SAR data. IEEE Trans. Geosci. Remote Sens. **31**(4), 896–906 (1993)
63. Schmitt, A., Wessel, B., Roth, A.: An innovative curvelet-only-based approach for automated change detection in multi-temporal SAR imagery. MDPI Remote Sens. **6**(3), 2435–2462 (2014)
64. Serpico, S., Bruzzone, L.: Change detection (Ch. 15). In: Chen, C.H. (ed.) Information Processing for Remote Sensing, World Scientific, Singapore (1999)

65. Singh, A.: Digital change detection techniques using remotely-sensed data. Int. J. Remote Sens. **10**(6), 989–1003 (1989)
66. Solano Correa, Y.T., Bovolo, F., Bruzzone, L.D: Change detection in very high resolution multisensor optical images. In: Proceedings of SPIE 9244, Image and Signal Processing for Remote Sensing XX, pp. 924,410–924,410 (2014)
67. Solano Correa, Y.T., Bovolo, F., Bruzzone, L.: VHR time-series generation by prediction and fusion of multi-sensor images. In: Proceedings of IEEE IGARSS'15, pp. 3298–3301 (2015)
68. Solano Correa, Y.T., Bovolo, F., Bruzzone, L.: An approach to multiple change detection in multisensor VHR optical images based on iterative clustering. In: Proceedings of IEEE IGARSS'16, pp. 5149–5152 (2016)
69. Solberg, A.H.S., Taxt, T., Jain, A.K.: A Markov random field model for classification of multisource satellite imagery. IEEE Trans. Geosci. Remote Sens. **34**(1), 100–113 (1996)
70. Vapnik, V.: The Nature of Statistical Learning Theory, 2nd edn. Springer, New York (2000)
71. Wang, F.: Fuzzy supervised classification of remote sensing images. IEEE Trans. Geosci. Remote Sens. **28**(2), 194–201 (1990)
72. Zanetti, M., Bovolo, F., Bruzzone, L.: Rayleigh-Rice mixture parameter estimation via EM algorithm for change detection in multispectral images. IEEE Trans. Image Process. **24**(12), 5004–5016 (2015)

Chapter 9
Satellite Image Time Series: Mathematical Models for Data Mining and Missing Data Restoration

Nicolas Méger, Edoardo Pasolli, Christophe Rigotti, Emmanuel Trouvé and Farid Melgani

Abstract One of the exceptional advantages of spaceborne remote sensors is their regular scanning of the Earth surface, resulting thus in Satellite Image Time Series (SITS), extremely useful to monitor natural or man-made phenomena on the ground. In this chapter, after providing a brief overview of the most recent methods proposed to process and/or analyze time series of remotely sensed data, we describe methods handling two issues: the unsupervised exploration of SITS and the reconstruction of multispectral images. In particular, we first present data mining methods for extracting spatiotemporal patterns in an unsupervised way and illustrate this approach on time series of displacement measurements derived from multitemporal InSAR images. Then we present two methods which aim to reconstruct multispectral images contaminated by the presence of clouds. The first one is based on a linear contextual prediction mode that reproduces the local spectro-temporal relationships characterizing a given time series of images. The second method tackles the image reconstruction problem within a compressive sensing formulation and with different implementation strategies. A rich set of illustrations on real and simulated examples is provided and discussed.

N. Méger · E. Trouvé
LISTIC laboratory, Polytech Annecy-Chambéry, Université Savoie Mont Blanc, BP 80439, 74944 Annecy cedex, France

E. Pasolli
Centre for Integrative Biology, University of Trento, Via Sommarive 9, 38123 Trento, Italy

C. Rigotti
INSA-Lyon, LIRIS, UMR5205, University Lyon, CNRS, INRIA, 20 av. A. Einstein, F-69621 Villeurbanne, France

F. Melgani (✉)
Department of Information Engineering and Computer Science, University of Trento, Via Sommarive 9, 38123 Trento, Italy
e-mail: melgani@disi.unitn.it

G. Moser and J. Zerubia (eds.), *Mathematical Models for Remote Sensing Image Processing*, Signals and Communication Technology,
https://doi.org/10.1007/978-3-319-66330-2_9

9.1 Introduction

In the last few years, Satellite Image Time Series (SITS) has gained particular interest in the remote sensing community as a relevant resource for Earth monitoring. Actual spaceborne remote sensors are characterized by growing spatial and spectral resolutions and permit a regular scanning of the Earth's surface, which is extremely useful to understand relationships between natural and man-made phenomena in order to promote better decision making. For example, in terms of optical sensors, the Taiwanese Formosat-2 satellite is already providing images with high-temporal resolutions, but with only four spectral bands and with a limited coverage of the Earth's surface. Additionally, the European Space Agency's (ESA) Sentinel-2 mission is producing global cover every five days with 13 spectral bands and with 10 to 60 m spatial resolution. Several different applications can take advantage from this kind of data, which can be especially utilized to monitor and detect changes over time. A summary of the most investigated applications based on SITS includes land-use and land-cover change mapping; forest and vegetation change mapping; forest mortality, defoliation, and damage assessment; deforestation, regeneration, and selective logging mapping; wetland change mapping; forest fire and fire-affected area detection; landscape change mapping; urban change mapping; environmental change mapping including monitoring of drought, flood, and coastal marine environments, desertification and landslide area detection; crop and shifting cultivation monitoring; road mapping. In order to efficiently analyze the typical large amount of data associated with SITS, appropriate processing methods have to be developed.

SITS acquired by active sensors usually comes from Synthetic Aperture Radar (SAR) operated on repeated orbits, which provides large regular time series thanks to the all-weather capability of this sensor: acquisitions are independent from the presence of clouds and sun illumination. As mentioned in Chap. 4, the amplitude and the phase of those complex images can be used to monitor the two main kind of "temporal evolution": changes on the land cover (abrupt or progressive changes) and surface displacements by offset tracking or by differential interferometry (D-inSAR) to measure up to millimeter deformations due for instance to seismic activity (see example in Sect. 9.2).

The main difficulties in the use of the amplitude information come from the speckle effect due to coherent imagery and the presence of distributed scatterers in the resolution cells. As discussed in Chaps. 4 and 5, this phenomenon is often modeled as a multiplicative noise with non-gaussian statistical distributions (Rayleigh, Gamma, Fisher, Nakagami, etc.) depending on the "number of looks" averaged to reduce this noise and the presence of texture on the ground [1]. One of the advantage of using SAR time series is the possibility to reduce this noise by multitemporal filtering in order to preserve as much as possible the spatial resolution. Different approaches have been developed [2], including 3D region growing techniques [3] and optimal noise reduction [4] or multiplicative wavelets along the temporal axes [5]. One of the drawback of such approaches is the risk of merging (blurring) the temporal information which is useful for change analysis. To avoid this effect, some

authors propose to combine change detection and multitemporal filtering techniques to give priority to temporal neighborhoods made of pixels belonging to the same statistical population [6, 7]. In most of those approaches, statistical models are used to describe the variability of the SAR amplitude, to measure the homogeneity of the populations, or to derive parameters [8] or metrics used to detect changes or to weight and aggregate pixels.

The phase information is also used either with polarimetric data to characterize the scattering mechanisms (surface, double bound, volume scattering, etc.) by the so called PolSAR decomposition techniques (see Chap. 5), or with interferometric data (see Chap. 4) to measure distance differences by the phase difference, or a combined use of those techniques (PolInSAR data) to retrieve more advanced information such as forest, crop, or snow/ice volume thanks to the radar penetration. Such applications also benefit from the use of multitemporal data set either to improve the estimation of physical parameters thanks to the use of spatiotemporal neighborhoods, or to discriminate the sought after signal (usually deterministic models) from the random behavior of some sources of uncertainty. In the case of SAR interferometry, the atmospheric artifacts due to the different meteorological conditions at different dates, the lack of precision in orbits and elevation models, and the phase unwrapping problem are the main sources of difficulty. Several multitemporal techniques have been developed to combine interferograms in order to minimize those error sources, with two main approaches. The permanent scatterers approach [9] consists in selecting only specific targets which are less affected by temporal and baseline decorrelation and to separate topographic errors from the displacement signal on the phase difference time series measured on those points. The small baseline subset (SBAS) approach consists in using only small temporal or spatial baseline to build a redundant network of interferograms, to correct and unwrap them in an iterative process which allows the temporal series of displacement to be inverted [10]. The use of mathematical inversion techniques combined with appropriate models is often the key issue to make those approaches successful.

The analysis of SITS acquired by optical sensors can be subdivided into three main groups, depending on the use of the time dimension [11–13]. In the first category, the time is used just as an attribute identifier, i.e., the different images are concatenated into a single data structure without taking into account the real order between the images. For example, some strategies are based on a linear transformation of the data such as principal component analysis or maximum auto-correlation factor [14]. Other methods are based on classification algorithms, which can be applied directly on the concatenated image or by classifying independently each image and then by combining the different classification maps [15]. The second category consists in using the temporal information as a partial ordering. These methods were initially proposed for bi-temporal images, but can be extended to multitemporal analysis by concatenating them. Basic strategies combine the values of the image at time t1 and those at time t2 by simple [16] or more advanced operators [17]. Other widely used methods are represented by change vector analysis [18] or linear regression [19], in which the values at time t1 are supposed to be linearly correlated with those at time t2. In the last category, the entire ordering associated with the temporal

information is considered. For example, it includes frequent pattern mining [20], in which frequent sub-sequences of radiometric evolutions are extracted, and frequency analysis [21], in which radiometric series are processed with Fourier or wavelet decompositions. Although the large amount of approaches proposed for the analysis of SITS, a common assumption is that the considered images are not affected by the presence of clouds. However, this represents a frequent problem for passive sensors, which are high sensitivity to weather conditions during the image acquisition process. The presence of clouds can be viewed as a source of contamination that makes the images partly or completely useless for assessing landscape properties. A solution to this issue is given in [13], in which the authors have proposed an approach able to deal with irregularly sampled SITS. The problem of reconstructing areas of the images contaminated by the presence of clouds will be analyzed in detail in the rest of the chapter.

In this chapter, two important issues related to the analysis of SITS are considered and handled by different approaches. The first issue deals with knowledge discovery in SITS and can be applied to any kind of physical or statistical parameter (time series of optical albedo, SAR radiometry, displacement fields, etc.) while the second issue deals with the reconstruction of missing data in multispectral images. In particular, we first present data mining methods for extracting spatiotemporal patterns in an unsupervised way and illustrate the potential of this approach on time series of displacement measurements derived from multitemporal InSAR images. For the reconstruction of multispectral images contaminated by the presence of clouds, two methods are presented: the first one is based on a linear contextual prediction model that reproduces the local spectro-temporal relationships characterizing a given time series of images, whereas the second one tackles the image reconstruction problem within a compressive sensing formulation and with different implementation strategies. A rich set of illustrations on real and simulated examples is provided and discussed.

9.2 Data Mining Methods for Spatiotemporal Pattern Extraction

9.2.1 Objectives and Originality of the Approach

Besides techniques for refining measurements and getting more precise estimates such as those presented in this book, a new kind of technique for processing *Satellite Image Time Series* (SITS) arises. These techniques are data mining ones: they sift through large data sets in an automated way to achieve *Knowledge Discovery in Databases (KDD)*. KDD is defined as "the non-trivial extraction of implicit, unknown, and potentially useful information from data" [22]. Used in a first a place, they produce descriptions of the data sets which can be further assessed in a more application-oriented way with dedicated tools such as ad-hoc statistical tests or clas-

sification systems. These descriptions can also be employed for SITS indexing and retrieval. A number of techniques have been proposed so far, such as searching for image/object/region evolutions (e.g., [23] or [24]) or directly extracting pixel-based spatiotemporal objects that are defined with respect to the temporal dimension (e.g., [25] or [26]) and also the spatial one (e.g., [27] or [28]). Complementarily to such extractions, we present data mining methods[1] for summarizing optical or radar SITS with the aim of assisting end users in browsing a SITS in a rapid, unsupervised, and human readable manner. The first one relies on an unsupervised data mining technique for finding pixel evolutions affecting a minimum number of pixels with sufficiently high spatial connectivity. Using these evolutions, namely *Grouped Frequent Sequential patterns (GFS-patterns)*, we build summaries of SITS by means of *SpatioTemporal Localization maps (STL-maps)*. These maps show the location of pixel evolutions in space and time, providing a characterization of the data set spatially and dynamically. The selection of a set of representative STL-maps is achieved using a second method exploiting a randomization procedure and a *Normalized Mutual Information (NMI)*-based scoring. In practice, it turns out to be effective in finding interesting groups of pixels, sharing common temporal evolutions, and that would not have been exhibited by other approaches. This part is organized as follows: Sect. 9.2.2 presents GFS-patterns while Sect. 9.2.3 details STL-maps and their NMI-based scoring procedure. Finally, Sect. 9.2.4 concludes this part of the chapter.

9.2.2 Grouped Frequent Sequential Patterns

In this section, a data mining pattern for exploring SITS is presented: the *Grouped Frequent Sequential Pattern (GFS-Pattern)*. Some preliminary definitions are given to define a SITS as a set of temporal sequences from which a common kind of data mining pattern, the sequential pattern, can be extracted. Then, the connectivity measure used to define the GFS-patterns is introduced and examples of GFS-patterns are given. Finally, the extraction of GFS-patterns is detailed.

9.2.2.1 Preliminary Definitions

Let us consider a SITS that covers the same area at different dates. Within each image, each pixel is associated with a value, for example, the reflectance intensity or the phase difference of the geographical zone that it represents. We can transform those pixel values into values belonging to a discrete domain, using labels for encoding pixel states. Those labels can correspond to ranges obtained by image quantization or to pixel classes resulting from an unsupervised classification (e.g., using K-means or EM-based clustering).

[1]The authors wish to thank the ANR EFIDIR and ANR FOSTER projects for funding the works presented in this chapter.

Definition 1 (*label and pixel state*) Let $L = \{i_1, i_2, \ldots, i_s\}$ be a set containing s distinct symbols called *labels*, used to encode the values associated with the pixels. A *pixel state* is an ordered pair (e, t) where $e \in L$ and $t \in \mathbb{N}$, such that t is the occurrence date of e. The date t is simply the time stamp of the image from which the value e has been obtained.

Subsequently, we define a *symbolic SITS* as a set of *pixel evolution sequences*, with each sequence describing the states of a pixel over time/at different dates.

Definition 2 (*pixel evolution sequence and symbolic SITS*) For a pixel p, the *pixel evolution sequence* is a pair $((x, y), seq)$, where (x, y) are the coordinates of p, and seq is a tuple of pixel states $seq = \langle (e_1, t_1), (e_2, t_2), \ldots, (e_n, t_n) \rangle$ containing the states of p ordered by increasing dates of occurrences. A *symbolic SITS* (or SITS when clear from the context) is then a set of pixel evolution sequences.

For a typical symbolic SITS, we thus get a set of millions of pixel evolution sequences, each sequence containing the discrete descriptions of the acquisition values of a given pixel.

9.2.2.2 Sequential Patterns

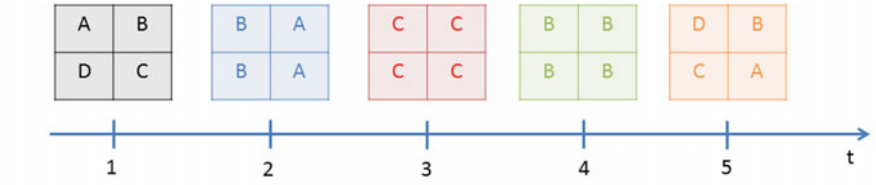
An important and active data mining research area is the mining of *bases of sequences* to extract *sequential patterns* [29]. This domain is now mature and provides efficient techniques for extracting such patterns. A typical base of sequences is a set of sequences of discrete events, in which each sequence has a unique sequence identifier. For SITS, if we consider the pairs (x,y) of coordinates of the pixels as identifiers of their evolution sequences, then a symbolic SITS is a base of sequences, and the standard notions [29] of sequential patterns and sequential pattern occurrences can be adapted as follows:

Definition 3 (*sequential pattern*) A *sequential pattern* α is a tuple $\langle \alpha_1, \alpha_2, \ldots, \alpha_m \rangle$ where $\alpha_1, \ldots, \alpha_m$ are labels in L, and m is the *length* of α. Such a pattern is also denoted as $\alpha_1 \rightarrow \alpha_2 \rightarrow \ldots \rightarrow \alpha_m$.

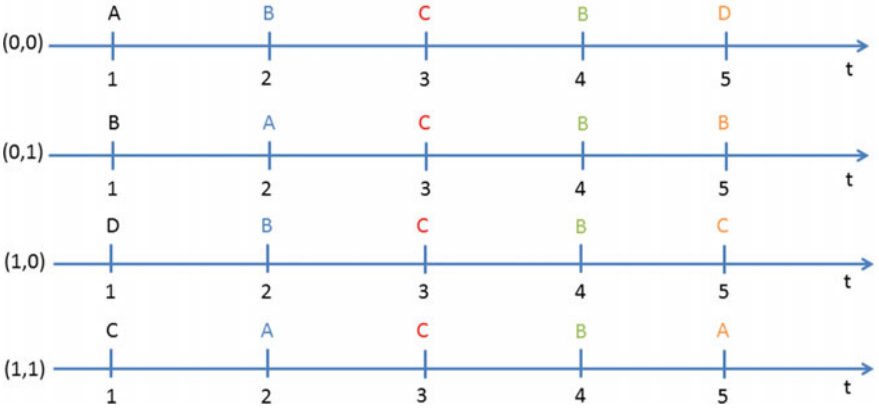
Definition 4 (*occurrence and support*) Let $\mathcal{S}$ be a symbolic SITS, and $\alpha = \alpha_1 \rightarrow \alpha_2 \rightarrow \ldots \rightarrow \alpha_m$ be a sequential pattern. Then, $((x, y), \langle (\alpha_1, t_1), (\alpha_2, t_2), \ldots, (\alpha_m, t_m) \rangle)$, where $t_1 < t_2 < \ldots < t_m$, is an *occurrence* of α in $\mathcal{S}$ if there exists $((x, y), seq) \in \mathcal{S}$ such that (α_i, t_i) appears in seq for all i in $\{1, \ldots, m\}$. Such a pixel evolution sequence $((x, y), seq)$ is said to support α. The *support* of α in $\mathcal{S}$, denoted by $support(\alpha)$, is simply the number of sequences in $\mathcal{S}$ that support α.

Example 1 A mock symbolic SITS containing the states of four pixels.

$$
\begin{aligned}
&((0, 0), \langle (A, 1), (B, 2), (C, 3), (B, 4), (D, 5) \rangle),\\
&((0, 1), \langle (B, 1), (A, 2), (C, 3), (B, 4), (B, 5) \rangle),\\
&((1, 0), \langle (D, 1), (B, 2), (C, 3), (B, 4), (C, 5) \rangle),\\
&((1, 1), \langle (C, 1), (A, 2), (C, 3), (B, 4), (A, 5) \rangle)
\end{aligned}
$$



(a) Representation of Example 1.



(b) Another representation of Example 1.

Fig. 9.1 Two equivalent representations of Example 1

This data set, as shown by Fig. 9.1, describes the evolution of four pixels through five images acquired at time 1, 2, 3, 4, and 5 and using $L = \{A, B, C, D\}$.

For example, the successive discrete labels associated with the values of the pixel located at $(0, 0)$ are A, B, C, B, and D. In this data set, the sequential pattern $A \rightarrow C \rightarrow B$ has the following four occurrences (notice that the elements in an occurrence do not need to be contiguous in time):

$$
\begin{aligned}
&((0, 0), \langle (A, 1), (C, 3), (B, 4)\rangle),\\
&((0, 1), \langle (A, 2), (C, 3), (B, 4)\rangle),\\
&((0, 1), \langle (A, 2), (C, 2), (B, 5)\rangle),\\
&((1, 1), \langle (A, 2), (C, 3), (B, 4)\rangle)
\end{aligned}
$$

The pattern has four occurrences but appears in only three different pixel evolution sequences: its support is $support(A \rightarrow C \rightarrow B) = 3$. Finally, it should be pointed out that a label can be repeated within a pattern: pattern $C \rightarrow C$ has two occurrences, one in the third and one in the fourth sequence.

The number of different patterns occurring in a data set can be high. Therefore, only the frequent ones are selected by using a support threshold.

Definition 5 (*frequent sequential pattern*) Let σ be a strictly positive integer termed as *support threshold*. Let α be a sequential pattern, then α is a *frequent sequential pattern* if $support(\alpha) \geq \sigma$. The support threshold can also be specified as a relative threshold $\sigma_{rel} \in [0, 1]$. Then, a pattern α is frequent if $support(\alpha)/|\mathcal{S}| \geq \sigma_{rel}$, where $\mathcal{S}$ is the data set and $|\mathcal{S}|$ is the number of sequences in $\mathcal{S}$.

Such a support constraint is used by sequential pattern extraction algorithms to reduce the search space and to achieve reasonable execution times.

9.2.2.3 Spatial Connectivity

Sequential patterns SITS analysis leads to a natural interpretation of the notion of support. For a pattern α, the support of α is simply an area, i.e., the total number of pixels in the image having an evolution in which α occurs. These pixels are said to be *covered* by α.

Definition 6 (*covered pixel*) A pixel having the evolution sequence $((x, y), seq)$ is *covered* by a sequential pattern α if α has at least one occurrence in seq. The set of the coordinates of the pixels covered by α is denoted by $cov(\alpha)$.

Therefore, for a frequent pattern α, the threshold σ (resp. σ_{rel}) can be interpreted as the minimum area (resp. relative area) that must be covered by α. However, a threshold on the covered area is not sufficient because, most of the time, interesting parts in images are made of pixels forming regions. An additional criterion, the *average connectivity* measure, is thus introduced. It is based on the *8-nearest neighbors (8-NN)* convention. Using this measure, the algorithm extracts patterns that cover pixels forming groups which can be defined as follows:

Definition 7 (*local connectivity*) For a symbolic SITS $\mathcal{S}$, let $occ((x, y), \alpha)$ be a function that, given the spatial coordinates (x, y) and a sequential pattern α, indicates whether α occurs in $\mathcal{S}$ at location (x, y). More precisely, $occ((x, y), \alpha)$ is equal to 1 if and only if there is a sequence seq in $\mathcal{S}$ at coordinates (x, y) and α occurs in $((x, y), seq)$. Otherwise $occ((x, y), \alpha)$ is equal to 0. If α occurs in $((x, y), seq)$, then its *local connectivity* at location (x, y) is $LC((x, y), \alpha) = [\sum_{i=-1}^{i=1} \sum_{j=-1}^{j=1} occ((x + i, y + j), \alpha)] - 1$.

The value $LC((x, y), \alpha)$ is simply the number of pixels in the 8-neighborhood of (x, y) having an evolution supporting α. The sum is decremented by one, so as not to count the occurrence of α at location (x, y).

Definition 8 (*average connectivity*) The *average connectivity* of α is defined as:

$$AC(\alpha) = \frac{\sum_{(x,y) \in cov(\alpha)} LC((x, y), \alpha)}{|cov(\alpha)|}$$

For the pixels supporting α, this measure gives the average number of neighbors in their 8-NN that also support α. In Example 1, $AC(A \rightarrow C \rightarrow B) = 6/3 = 2$ and $AC(C \rightarrow C) = 2/2 = 1$.

In addition to the average connectivity, the notion of *super-pattern* is also required to define *grouped frequent sequential patterns*. For the simple form of sequential patterns used in this book, the notion of *super-patterns* can be defined as follows.

Definition 9 (*super-pattern*) A sequential pattern $\beta = \beta_1 \rightarrow \beta_2 \rightarrow \ldots \rightarrow \beta_m$ is a *super-pattern* of a sequential pattern $\alpha = \alpha_1 \rightarrow \alpha_2 \rightarrow \ldots \rightarrow \alpha_n$ if $m > n$ and if there exist integers $1 \leq i_1 < i_2 < \ldots < i_n \leq m$ such that $\alpha_1 = \beta_{i_1}$, $\alpha_2 = \beta_{i_2}$, …, $\alpha_n = \beta_{i_n}$.

Finally, *grouped frequent sequential patterns* are defined as follows.

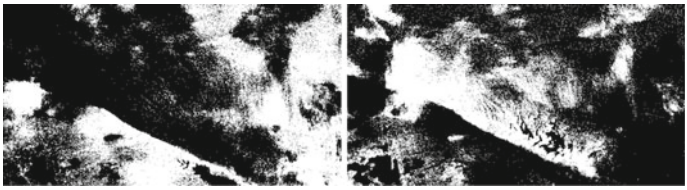
Definition 10 (*grouped frequent sequential pattern (GFS-patterns)*) Let $\mathcal{S}$ be a symbolic SITS. Given a sequential pattern α frequent in $\mathcal{S}$, and a positive real number κ termed *average connectivity threshold*, α is said to be a *Grouped Frequent Sequential pattern* (*GFS-pattern*) if $AC(\alpha) \geq \kappa$ in $\mathcal{S}$ and if $\nexists \beta$ such that β is a super-pattern of α, with β a frequent sequential pattern with $AC(\beta) \geq \kappa$.

For instance, in Example 1, if $\sigma = 2$ and if $\kappa = 2$, then $A \rightarrow C \rightarrow B$ is a GFS-pattern while $C \rightarrow C$ is not: its average connectivity measure does not exceed κ. Though $A \rightarrow B$ could be retained with respect to σ and κ, it is not a GFS-pattern: $A \rightarrow C \rightarrow B$ is a super-pattern of $A \rightarrow B$. This *maximality* constraint makes it possible to focus on the most specific patterns [30].

The concept of GFS-pattern has been defined in [31] without any maximality constraint. Nevertheless, when assessing the results from qualitative point of view, in [31] and in all following papers, we mainly focused on patterns that are maximal ones. Subsequently, in this book, we include this constraint into the definition of GFS-patterns themselves.

Finally, it is worth noting that:

- the different pixel states of the occurrence of a GFS-pattern are not necessarily consecutive,
- pixels sharing a same pattern do not need to be synchronized in time,
- no timing constraint is set,
- the shape of the observed phenomena is not set beforehand,
- a wide range of scales can be taken into account (all surfaces greater or equal to σ),
- pixels sharing a same pattern form objects that are coherent spatially (average connectivity greater or equal to κ),
- a same pixel can be covered by zero, one or more GFS-patterns (GFS-pattern extraction is not another segmentation technique).



(a) Pattern #1. (b) Pattern #2.

Fig. 9.2 Localization (*white pixels*) of 2 GFS-patterns. Pixels are **a** getting closer to the satellite and **b** away from the satellite

9.2.2.4 Examples of GFS-Patterns

In [31] and in [32], experiments on optical SITS show the potential of GFS-patterns for applications such as agriculture monitoring. As reported in [31] or in [33], GFS-patterns can also be extracted from SAR SITS. For example, in [33], an InSAR time series of 24 images (701×701 pixels) containing the cumulated phase evolution/displacement for each acquisition dates was mined. Both ground deformation and atmospheric turbulences contribute to the phase evolution/displacement[2]. The series was built from 25 *Environmental Satellites (ENVISAT)*[3]; SAR images acquired over the 2004–2009 period and covering the Haiyuan seismic fault in the north-eastern boundary of the Tibetan plateau (about 50 km $\times$ 50 km). Once again, the quantization was done using the 33rd and the 66th percentiles and threshold κ was set to 6. Since end users are interested in phenomena covering large areas, threshold σ was then set to 20%.

Among the extracted GFS-patterns, the two following ones were reported:

- pat. #1: $1 \rightarrow 1 \rightarrow 1 \rightarrow 1 \rightarrow 1 \rightarrow 1 \rightarrow 1 \rightarrow 1 \rightarrow 1 \rightarrow 1$;
- pat. #2: $3 \rightarrow 3 \rightarrow 3 \rightarrow 3 \rightarrow 3 \rightarrow 3 \rightarrow 3 \rightarrow 3 \rightarrow 3 \rightarrow 3$;

Pattern #1 indicates that some areas tend to get closer to the satellite (symbol "1") while pattern #2 shows that other areas are getting away from the satellite (symbol "3"). With regard to symbol "2", it is associated with very weak displacements. As it can be observed in Fig. 9.2 which focuses on the center part of the zone, these patterns are spatially complementary.

A creep phenomenon is thus revealed by these two patterns. It is coherent with the motion of the northern part of the studied zone (upper part of the image) that has been reported by the experts. Furthermore, the localization of the seismic fault that is due to this creep phenomenon can be inferred by looking at the crisp frontier

[2]We thank Romain Jolivet (California Institute of Technology-CALTECH, Seismological laboratory) and Cécile Lasserre (Centre National de la Recherche Scientifique-CNRS, ISTerre laboratory) for computing and providing this InSAR time series.

[3]We thank the European Space Agency (ESA) for providing this ENVISAT SAR series (Dragon project ID5305).

between affected and non-affected pixels, especially on the upper-left to lower-right diagonal of the images.[4]

All of these results show that GFS-patterns are quite general and can be extracted from different types of SITS (optical data, radar data, different resolutions). Applications range from agricultural monitoring to military surveillance or to ground deformation monitoring. Though no advanced preprocessing such as cloud masking was performed, it should be noted that GFS-patterns offer end users the possibility of exploring huge time series and discovering temporal evolutions which could be hidden by random uncertainty such as atmospheric turbulences. In addition, no GFS-pattern expressing such phenomena was extracted during the various experiments we reported so far. Indeed, the conjunction of the different constraints (minimum surface, minimum average connectivity, maximality) discards phenomena that are dispersed over time and space. Finally, all the experiments reported so far show that GFS-patterns can be extracted on standard PCs (e.g., Intel Core 2 @3 GHz, 4 GB RAM, Linux kernel 2.6) within reasonable amount of time. This can be achieved by pushing the constraints set on the support and the average connectivity in the extraction process, which leads to an efficient pruning of the search space. The reader is referred to [32] for more details.

9.2.3 STL-Maps and NMI-Based Scoring of STL-Maps

This section presents a method for assessing GFS-patterns. It relies on SpatioTemporal Localization maps (STL-maps). These maps show the location of GFS-patterns occurrences in space and time, providing a characterization of the data set spatially and dynamically. They are defined in Sect. 9.2.3.1. The most representative STL-maps, and thus GFS-patterns, are selected using a randomization procedure and a scoring based on the Normalized Mutual Information (NMI) measure which are detailed in Sect. 9.2.3.2.

9.2.3.1 SpatioTemporal Localization Maps (STL-Maps)

Once a GFS-pattern has been extracted, a convenient way to assess it is to build an image where all pixels are set to black except those that are covered by the pattern. This localization gives the spatial information at a glance while the temporal evolution is given by the pattern itself. For non-black pixels, a refinement consists of using several colors giving access to a complementary temporal information such as the time spans or the ending dates of the pattern occurrences. In order to check

[4]We thank Romain Jolivet (California Institute of Technology-CALTECH, Seismological laboratory), Cécile Lasserre (Centre National de la Recherche Scientifique-CNRS, ISTerre laboratory), and Catherine Pothier (Institut National des Sciences Appliquées de Lyon-INSA Lyon, LGCIE laboratory) for this interpretation.

whether a pattern propagates or not in space, we propose to use a color scale associated with the ending dates (or ending image numbers) of the earliest pattern occurrences. As reported in [33, 34], such an image of dates brings very useful information as it summarizes the occurrences of a pattern, spatially and dynamically. If the ending date are linearly linked to the color palette shown in Fig. 9.3d, the spatial localizations presented in Fig. 9.2 can be refined into the STL-maps given by Fig. 9.3a, b. As it can be observed, pattern #1 and pattern #2 propagate in time and space. The propagation of pattern #1 is not radial w.r.t. to the creeping zone, and creep migration along the fault could explain such patterns.[5]

9.2.3.2 Swap Randomization of SITS and NMI-Based Scoring of STL-Maps

Each STL-map being related to a single GFS-pattern, depending on the SITS, one may obtain numerous STL-maps. For example, in Sect. 9.2.2.4, 1673 GFS-patterns were extracted from the SAR SITS. In order to draw attention to the most promising patterns, we propose to select a set of representative STL-maps. To this end, we propose to target two kinds of STL-maps. Either they will be required to convey more information than what could be expected when considering a randomized SITS with the same structure in terms of symbols frequencies, or they should contain the most prominent phenomena with respect to the spatiotemporal distribution of the symbols. The SITS randomization is achieved by a swap randomization technique adapted from a Boolean matrix randomization one. It is presented in Sect. 9.2.3.2. The comparison between the STL-maps obtained on the original and the randomized SITS relies on the Normalized Mutual Information (NMI). The latter will be used as a measure to score STL-maps and produce an NMI-based ranking of STL-maps. This ranking will make possible to draw attention to the two kinds of STL-maps that are targeted.

Swap Randomization of SITS

Randomization is aimed at hypothesis testing and numerous works such as [35] shows their importance. With regard to swap randomization as proposed in [36] or [37], it is basically applied to Boolean matrices. The bottom line is to compare results obtained for a given data set against the results obtained for a set of randomized data sets having the same structure in terms of row and column margins (sums). In other words, the information conveyed by these margins is not what is studied: it is the arrangement of the data set itself, which can be expressed by patterns such as rectangles of 1's. In [38], we adapted the swap randomization of Boolean matrices to the swap randomization of symbolic matrices and showed that randomized data sets are still equiprobable. Let S be the original SITS and S' the randomized one. S' is generated from S by

[5] We thank Romain Jolivet (California Institute of Technology-CALTECH, Seismological laboratory), Cécile Lasserre (Centre National de la Recherche Scientifique-CNRS, ISTerre laboratory) and Catherine Pothier (Institut National des Sciences Appliquées de Lyon-INSA Lyon, LGCIE laboratory) for this interpretation.

applying a series of symbol swaps, each swap being applied to the latest matrix that has been obtained. Each swap is defined as follows: let us consider a SITS as a $m \times n$ symbolic matrix where each row denotes a pixel position and where each column corresponds to an acquisition date. Each cell of the matrix is thus a pixel state. Pixel states are described using the symbols defined by end users (cf. Sect. 9.2.2.1). Let D be such a symbolic matrix. Let u and v be two pixel positions chosen at random. Let i and j be two dates also chosen randomly. If $D_{u,i} = D_{v,j} = \alpha$ and $D_{u,j} = D_{v,i} = \beta$ with α and β two distinct symbols describing the latter pixel states, then pixel states are changed so that $D_{u,i} = D_{v,j} = \beta$ and $D_{u,j} = D_{v,i} = \alpha$: symbols α and β are swapped. When dealing with SITS, this kind of swap can be interpreted as a spatiotemporal swap. By construction, as for Boolean matrices, symbol frequencies are maintained both temporally (columns) and spatially (rows). In other words, an acquisition expressing the presence of vegetation is not transformed into an image of a desert. Similarly, a pixel evolution showing variations between snow and rocks will not be transformed into a urban sprawl pattern. The spatiotemporal nature of the acquisitions is maintained. Following the swap randomization procedure of Boolean matrices established in [36] or [37]:

- all pixel positions and all dates have the same probability of being chosen and can be chosen more than once,
- swaps are not final and can be undone by other swaps.

In practice, and so as to adopt conservative settings, the minimum number of swaps to be applied is set to be in the order of 10 times the number of pixel states that are contained in the SITS to be randomized.

NMI-Based Scoring of STL-Maps

How to compare the STL-map C, obtained on the original SITS for a pattern P, with C', the STL-map obtained for P on the swap-randomized SITS? At this stage, it should be recalled that we are interested by the following two settings:

- C and C' share little information: the informational content of C is singular as it can not be obtained for a randomized data set with the same structure in terms of symbols frequencies,
- C and C' share a lot of information: the swap randomization does not destroy the occurrences of P, which means that it expresses a prominent phenomenon.

Another question arises: would it be possible to distinguish the latter two settings using a single measure without making any assumption about the relation between C and C'?

Let Ω be the sample space containing all ending dates. Let us consider each ending date x of C as the realization of a discrete random variable X and each ending date

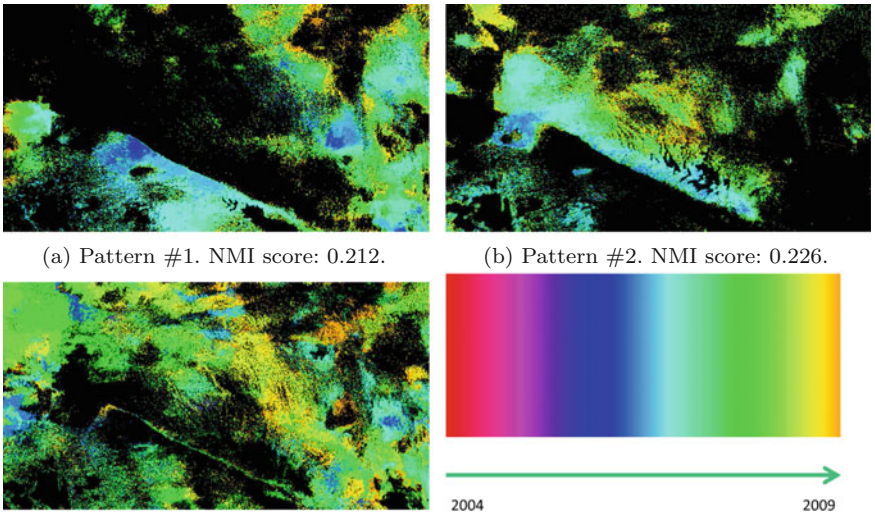
y of C' as the realization of a discrete random variable Y. In order to assess the information content shared by X and Y with values in the range[0; 1], we proposed in [38] to use a normalized version of the *mutual information* as defined in [39], the *Normalized Mutual Information (NMI)*:

$$NMI(X;Y) = \frac{\sum_{x,y\in\Omega^2} P(x,y)\log\frac{P(x,y)}{P(x)P(y)}}{min(H(X),H(Y))} \tag{9.1}$$

where $P(x, y)$ represents the probability of co-occurrence of the two ending dates x and y at the same pixel position, in C and C'. The NMI quantifies the information content shared by two random variables. In other words, knowing the realizations of two random variables X and Y, it measures the extent to which the realizations of variable X can be deduced from the ones of Y, and vice versa. Therefore, it can been seen as a measure of the mutual dependance of X and Y. By relying on the NMI, no assumption about the relation between the ending dates is done. Instead of using, for example, a Pearson correlation coefficient to check whether there exists a linear relation, the ending dates are just evaluated as labels by considering their co-occurrences. In addition to GFS-patterns extraction, this allows us to produce summaries which are as unsupervised as possible. Once the NMI is computed for X and Y, then C is scored using this quantity. Scoring is done for every STL-map which is thus ranked with other STL-maps according to its NMI-based score. As a result, each STL-map being related to a single GFS-pattern, GFS-patterns are also ranked. The NMI-based rankings that are obtained can be easily browsed. If one is interested by STL-maps/GFS-patterns showing phenomena that cannot be obtained on a swap-randomized SITS, then one has to look to STL-maps, and thus GFS-patterns, with low NMI scores. Conversely, if one is interested by STL-maps/GFS-patterns showing prominent phenomena that are still present in a swap-randomized SITS, then one has to consider STL-maps with high NMI scores. In order to build a summary containing the most representative STL-maps, both ends of the rankings have to be considered by picking up the STL-maps with the lowest and the highest NMI scores. Beside building up a summary, it has been shown in [38] that using the NMI and focusing on the ends of the ranking allows to consider a single randomized data set, as opposed to standard swap randomization methods that require thousands of randomized data sets to be generated. The implementation of the whole method has been done in Python 2.7 and in C, using our own data structures. All the experiments reported in [38] show that GFS-patterns can be extracted on standard PCs (e.g., Intel Core 2 @3 GHz, 4 GB RAM, Linux kernel 2.6) within reasonable amounts of time.

STL-maps and ranking examples

Back to Sect. 9.2.2.4 and the results obtained for the SAR SITS, we ranked the STL-maps of the 1673 extracted GFS-patterns by randomizing the SITS with 10, 000, 000 swaps. The average NMI score is 0.03 with a very weak standard deviation of 0.009. The maximum NMI score is 0.226 and the minimum one is 0.021. Among the 20 most prominent phenomena/STL-maps/GFS-patterns to be reported, i.e., the STL-



(a) Pattern #1. NMI score: 0.212. (b) Pattern #2. NMI score: 0.226.

(c) Pattern #3. NMI score: 0.024. (d) Color palette.

Fig. 9.3 STL-maps of 3 GFS-patterns built using the color palette shown in **d**. Pixels are **a** getting closer to the satellite, **b** away from the satellite, **c** away and then closer to the satellite (colour figure online)

maps having the highest NMI scores, pattern #1 and pattern #2 are found. Their STL-maps as well as their NMI scores are presented in Fig. 9.3a, b.

If end users are interested by rare phenomena, i.e., those that are destroyed by swap randomization, then STL-maps with low NMI values are to be examined. Among the 20 most rare phenomena, GFS-pattern #3: $2 \rightarrow 3 \rightarrow 2 \rightarrow 2 \rightarrow 1 \rightarrow 2 \rightarrow 2$ is of interest. As it can be seen in Fig. 9.3c, its STL-map directly shows the localization of the seismic fault that was inferred with pattern #1 and pattern #2 (upper-left to lower-right diagonal). These results show that STL-maps and their rankings can produce meaningful and useful summaries of SITS. Similar results on radar SITS have been reported in [40]. Encouraging results on optical SITS have been also reported in [41] and in [38].

9.2.4 Discussion

The unsupervised data mining technique presented in this section of the chapter can be summarized as shown in Fig. 9.4: starting from a symbolic SITS that can be built without any advanced preprocessing (no cloud masking, no calibration, etc.); GFS-patterns are extracted using only two intuitive parameters: σ, the minimum surface/support threshold and k, the minimum average connectivity. A randomized symbolic SITS is also generated from the same symbolic SITS. Only one parameter

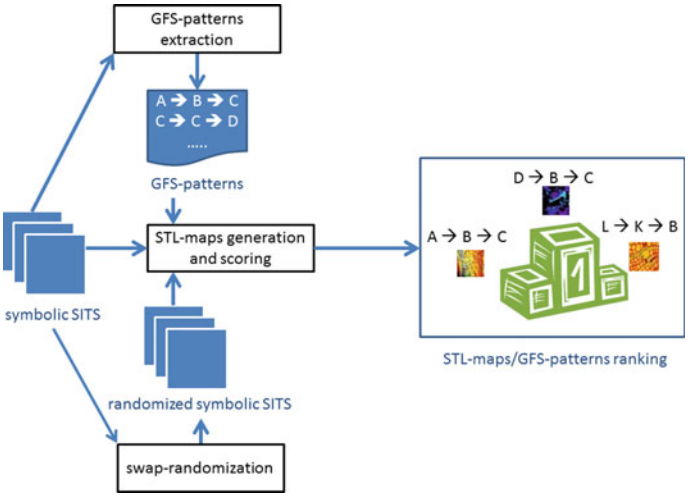


Fig. 9.4 GFS-patterns extraction and assessment

is required: the number of swaps, which can be inferred from the amount of pixel states. Then, the STL-maps of the extracted GFS-patterns are generated for both the symbolic SITS and the randomized one. Finally, the latter are used to compute the NMI for each STL-map/GFS-pattern. According to the scores that are obtained, a STL-maps/GFS-patterns ranking is produced. It makes it possible to browse a SITS in a human readable manner by having a closer look at prominent phenomena as well as rare ones. All the data mining methods presented in this section are made available for free through the *SITS-P2MINER* prototype [42]. It is written in C and in Python 2.7 and can be run on x64 platforms (Mac OS, Linux, Windows).

9.3 Reconstruction Methods for Multispectral Images

9.3.1 Survey

One of the major limitations of passive sensors is their high sensitivity to weather conditions during the image acquisition process. The resulting images are frequently subject to the presence of clouds, whose extent depends on the season and the geographic position of the study region. For instance, in Canada, from 50 to 80% of the Earth's surface can be obscured by clouds in mid-morning [43]. Depending on the application and the end-user requirements, clouds can be viewed: as a source of information for measuring important parameters such as cloud liquid water useful in meteorological forecasting and hydrological studies [44, 45]; or as a source of contamination that makes the image partly useless for assessing landscape properties. In the latter case, which represents the focus of this chapter, clouds distort the spectral response of land covers, thereby resulting in missing data for high-frequency passive

sensors including multispectral optical sensors and microwave radiometers. Several methodologies have been developed in the past in order to cope with this problem. Generally, the common approach first detects the contaminated regions and in a second instance, it attempts to remove the clouds by substituting them with cloud-free estimations.

Cloud detection is generally based on the assumption that clouds are colder (i.e., they emit less infrared radiation to space) and brighter (i.e., they reflect more solar radiation to space) than the ground. Other assumptions like their spatiotemporal variation in the visible and infrared ranges, which is higher than that of the surface, can also be exploited in the cloud detection process. Cloud detection consists of discriminating between clouds and the surface. In the literature, several approaches have been proposed to carry out this task, which has been referred to in many ways, including cloud masking and cloud screening. The common characteristic of cloud-masking approaches is that they reduce the cloud–noncloud classification problem to a problem of defining multiple thresholds optimized for the selected spectral bands [46–50].

The cloud removal problem, which is the focus of this part of the chapter, can be viewed as an image reconstruction/restoration issue, in which it is aimed at recovering an original scene from degraded or missing observations. Image reconstruction/restoration has been intensively and extensively studied in various application fields, such as radio astronomy, biomedical engineering, and machine vision, because of its practical importance as well as theoretical interest [51, 52]. In the remote sensing field, significant attention has been devoted to the reconstruction/restoration of images subject to various problems, such as acquisition blur and geometric distortions [53], phase distortions [54], resampling problems [55], or problems related to applications like buried object detection [56].

By contrast, less attention has been paid to the specific problem of cloud removal. Among the relatively few works available in the literature, one can find the simple image compositing technique, which consists of selecting the best measurement (i.e., the most cloud-free pixel) among a set of measurements acquired over a limited time period to represent the considered multitemporal pixel over that time period [57]. The main drawbacks of this technique are three: (1) it requires a high-temporal resolution acquisition over a short-time period; (2) it loses temporal resolution over the considered time period; and (3) it does not guarantee a cloud-free result since some areas may be cloudy on the whole image sequence. An interesting alternative adaptive reconstruction system was proposed in [58]. The authors assume that the temporal signature of a given pixel is contaminated by residual effects caused by imperfect sensing of the target and by spatially autocorrelated noise due to atmospheric attenuation. The system combines four different filters in an iterative way: a least-squares linear predictor for estimating missing or corrupted data at a specified time from previous history, two filters for determining, respectively, the spatial parameters and the mean intensity modeled by a Gibbs random field, required by the fourth filter to carry out a Bayesian reconstruction of the original intensity image. Though particularly effective, the system exhibits a high computational complexity and is not easily applicable to the general case of non-stationary temporal image series.

In [59], in order to circumvent the non-stationarity problem, the same authors propose to improve their adaptive system by substituting the linear predictor with an adaptive polynomial filter to track a better trend in the mean intensity process. In [60], another method is presented for recovering Advanced Very High Resolution Radiometer (AVHRR) measurements that are modified by the effects not only of clouds but also of cloud shadows. The method is based on the idea of detecting the contaminated pixels by analyzing their temporal NDVI profile and substituting them by interpolated values of individual channels or channel transformations. The method is simple and effective but presents the drawback of being limited to data acquired over vegetated areas. Some algorithms, like the second highest (SH) [61] and the modified maximum average (MMA) [62] have been developed specifically to remove cloud effects from Special Sensor Microwave/Imager (SSM/I) images. Both algorithms aim to produce a composite cloud-free image from a sequence of SSM/I images acquired over a short-time period. While the former is based on the idea of representing each image pixel by the second-highest value in the considered vector of multitemporal measurements as an alternative to the mean or median values, the latter removes cloud noise by averaging only part of the measurements contained in the vector. Selection of the MMA subset of measurements is carried out by considering all the measurements above the vector mean except the one with the highest value. In [62], the authors show that a hybrid algorithm that implements MMA in the presence of clouds and averages the measurements in their absence can significantly improve the quality of the composite image compared to the MMA and the SH algorithms. In [63], a mathematically well-founded method is proposed to remove the distortions caused by a particular kind of clouds from visible channels, namely cirrus clouds usually found above the 10 km altitude. The authors base their method on the fact that the measurements acquired at the 1.38 μm band are essentially due to cirrus reflectance attenuated by the absorption of water vapor contained in the uppermost layer of the atmosphere (above cirrus clouds) and develop a mathematical model for correcting those attenuation effects. They exploit this result to derive and then remove the true cirrus cloud reflectance from contaminated measurements in visible channels. This interesting method was assessed successfully on data acquired by two different hyperspectral sensors: the Moderate Resolution Imaging Spectro-Radiometer (MODIS) and the Airborne Visible Infrared Imaging Spectrometer (AVIRIS) sensors. In [64], an ecosystem classification-dependent temporal interpolation technique is proposed for reconstructing surface reflectance for MODIS data. It is based on the computation of pixel-level and regional ecosystem-dependent phenological curves. Missing temporal data associated with a given pixel are reconstructed from the most appropriate curve among the available pixel-level and regional curves. In [65], the authors propose a method to correct radiometric inconsistencies of cloud-contaminated images and their corresponding temporal images by generating a cloud-free mosaic image for a multitemporal SPOT data set. In order to ameliorate the transition between two mosaic parts, a wavelet-based fusion is adopted. More recently, a cloud removal method based on information cloning was developed [66]. The authors propose to clone cloud-free information from a set of multitemporal images, adopting a patch-

based reconstruction method formulated as a Poisson equation and solved using a global optimization process.

The main drawbacks that can be identified from the above-described cloud removal algorithms are: sensor-dependence, very-high-temporal resolution, ground cover-type dependence, cloud-type dependence, high methodological complexity, or/and limitation to composite image generation (i.e., incapability to reconstruct each image of a sequence separately). Two alternative approaches (based on linear contextual prediction [67] and compressive sensing [68], respectively) that aim to circumvent most of these drawbacks are presented in the following subsections of the chapter.

9.3.2 Problem Formulation

We consider a set of multitemporal multispectral images $I^{(i)}$ acquired over the same geographical area by an optical sensor at times t_i (with $i \in S = \{1, 2, , T\}$). Let us suppose that the images have been registered. We assume that (1) the images of the sequence may convey changes in the spectral appearance of objects on the ground and (2) they are characterized by an almost similar spatial structure. The last assumption can be considered realistic if the acquisition dates are close to each other (i.e., high-temporal resolution) or if the spatial dynamics of the geographical area under analysis is slow compared to the total time interval of the sequence (e.g., forest, mountainous, and urban areas). Moreover, we assume that the images have first been processed to generate a sequence of cloud/noncloud classification maps $M^{(i)}$ ($i \in S$) by using an automatic cloud-masking method or simply by photo interpretation. Given $M^{(i)}$, cloudy and non cloudy areas are represented by $\Omega^{(i)}$ and $\Phi^{(i)}$, respectively, subject to $I^{(i)} = \Omega^{(i)} \cup \Phi^{(i)}$. The specific problem of the detection of clouds (and their shadow) is not dealt with in this chapter. The objective of the investigated methods is to reconstruct any area contaminated by clouds (or by cloud shadows) for each image of the sequence. Therefore, each classification map $M^{(i)}$ will be used to guide the cloud removal process.

For the first investigated method based on linear contextual prediction, image channels are processed separately. Let us denote by $X^{(i)}$ ($i \in S$) one of the available sequences of T single-channel temporal images. We denote by C, a cloudy area of the image $X^{(i)}$.

The reconstruction problem of C in $X^{(i)}$ can be expressed as a problem of generating an image $Y^{(i)}$ such that:

$$Y^{(i)}(u, v) = \begin{cases} X^{(i)}(u, v) \text{ if } (u, v) \notin C \\ f\left[X^{(k)}(u, v)\right], k \in S_C, \text{ otherwise} \end{cases} \tag{9.2}$$

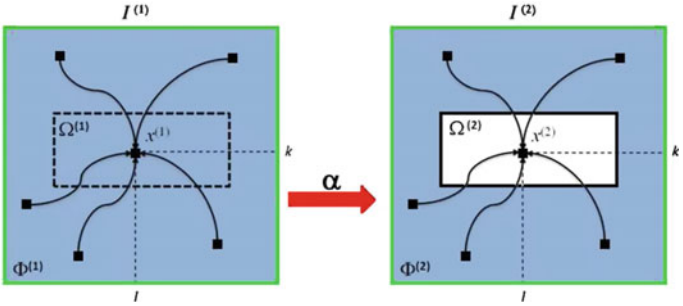


Fig. 9.5 Illustration of the reconstruction principle under a CS perspective

where (u, v) denotes pixel coordinates, $f[\cdot]$ represents a contextual prediction function, and S_c stands for the subset of indices corresponding to images $X^{(k)}$ $(k \neq i)$ that are cloud-free in the spatial area $N(C)$ including C and its neighborhood.

In other words, any image of the sequence obscured by the presence of a cloud in $N(C)$ will not contribute to the reconstruction process of C. In each image $X^{(k)}$ $(k \in S_c)$, the spatial area $N(C)$ can be subdivided into two cloud-free areas C and $\overline{C}$ such that $N(C) = C \cup \overline{C}$ and $C \cap \overline{C} = \emptyset$, where C represents the spatial area that corresponds to the cloudy area in $X^{(i)}$ and $\overline{C}$ stands for the neighboring spatial area.

For the second strategy based on compressive sensing, we simplify the problem by supposing to have just two images ($T = 2$) and we relax the above channel-based processing constraint (all channels are processed together). At this level, we make the hypothesis that the image $I^{(2)}$ has clouds, while the image $I^{(1)}$ is cloud-free. We assume that any pixel $x^{(1)} \in \Omega^{(1)}$ can be expressed as a linear combination of pixels in region $\Omega^{(1)}$ of $I^{(1)}$ (see Fig. 9.5).

In other words, in $I^{(1)}$, we have:

$$x^{(1)} = \Phi^{(1)} \cdot \alpha, \forall x^{(1)} \in \Omega^{(1)} \tag{9.3}$$

where α is an unknown weight vector associated with the considered pixel $x^{(1)}$ and having the same dimension as the number of pixels belonging to $\Phi^{(1)}$. The problem at this point is to infer $\alpha = f\left(\Phi^{(1)}, x^{(1)}\right)$.

Once α is computed, if we assume that $I^{(1)}$ and $I^{(2)}$ are temporally close, it will be possible to reuse the α coefficients to reconstruct the spatially corresponding pixel in the missing area $\Omega^{(2)}$, adopting the previous formulation for $I^{(2)}$, i.e., $\hat{x}^{(2)} = \Omega^{(2)} \cdot \alpha$ (see Fig. 9.5). In other words, for each pixel $x^{(1)} \in \Omega^{(2)}$, we evaluate α, and in a second moment, we reuse this weight vector to return an estimation of $x^{(2)} \in \Omega^{(2)}$:

$$\begin{aligned} \text{From } I^{(1)} : \alpha &= f(\Phi^{(1)}, x^{(1)}) \\ \text{To } I^{(2)} : \hat{x}^{(2)} &= \Phi^{(2)} \cdot \alpha \end{aligned} \tag{9.4}$$

where $f(\cdot)$ represents an estimation function. We recall that, differently from the first method, all image channels are processed simultaneously.

9.3.3 Linear Contextual Prediction Method

In the first investigated reconstruction method, given a contaminated image of the sequence, each area of missing measurements (i.e., cloudy or shadowed area) is recovered by means of a contextual prediction process that reproduces the local spectro-temporal relationships. These are deduced from the cloud-free areas in the spatial neighborhood of the contaminated region over the available series of temporal images. The contextual prediction process is carried out in two steps. First, a prediction system is trained to learn over $\overline{C}$; the temporal relationships between the set of available images $X^{(k)}$ $(k \in S_c)$ that are cloud-free in $N(C)$ on the one hand and the image $X^{(i)}$ on the other. This is done by implementing an ensemble of linear predictors, each trained in an unsupervised way over a local temporal region spectrally homogeneous in each temporal image of the sequence. In order to obtain such regions, each temporal image is locally classified in an unsupervised way by the Expectation–Maximization (EM) algorithm [69, 70] (already mentioned in the previous chapters with regard to various estimation tasks) assuming that the data (natural) classes are Gaussian. The number of data classes is estimated automatically by minimizing the Minimum Descriptive Length (MDL) criterion [71]. Once the training is completed, the prediction system is used to provide an estimate of each contaminated pixel of image $X^{(i)}$ in C, based on the spatially corresponding pixel values in images $X^{(k)}$ $(k \in S_c)$.

9.3.3.1 Contextual Prediction Process

The complexity of the relationship between images $X^{(k)}$ $(k \in S_c)$ and image $X^{(i)}$ in $N(C)$ will depend mainly on the complexity of their statistical distribution, which is conditioned by the quantity and quality of ground-cover classes in $N(C)$ at each date t_k $(k \in S_c \cup \{i\})$. In multispectral imagery, the assumption that the distribution of images can be approximated as a mixture of normally distributed samples is generally well-accepted. Accordingly, the probability distribution function (pdf) of each image $X^{(k)}$ $(k \in S_c)$ in $N(C)$ can be written as:

$$p_k(x) = \sum_{m=1}^{M_k} P(\omega_m^k) \cdot p(x|\omega_m^k) \tag{9.5}$$

where $P(\omega_m^k) = P_m^k$ and $p(x|\omega_m^k) = N(\mu_m^k, \sigma_m^k)$ are the prior probability and the conditional pdf associated with the mth gaussian mode in the $N(C)$ region of the kth image, respectively. Constant M_k stands for the number of modes characterizing the related pdf $p_k(x)$, while μ_m^k and σ_m^k are mean and standard deviation parameters, respectively. It is worth mentioning that a ground-cover class can be made up of more than one mode, each representing a spectral class of the data.

Given a multitemporal pixel vector $x = [x_1, x_2, , x_K]$ (K is the cardinality of S_c), such that x_j represents the pixel value in the jth image of the temporal sequence $X^{(k)}[k \in S_c; k = p(j)$, where $p(\cdot)$ is a mapping of the integers $\{1, 2, \ldots, K\}$ into $S_c]$ and $x_j \in \omega_{n_j}^j$ ($n_j \in \{1, 2, \ldots, M_j\}$), the contextual prediction function $f[\cdot]$ can be expressed as follows:

$$y = f[x] = \tilde{f}[x_1, x_2, \ldots, x_K | x_1 \in \omega_{n1}^1, x_2 \in \omega_{n2}^2, \ldots, x_K \in \omega_{nK}^K] \quad (9.6)$$

where $\widetilde{f}[\cdot]$ is a multitemporal mapping associated with the combination of modes $(\omega_{n1}^1, \omega_{n2}^2, \ldots, \omega_{nK}^K)$. Accordingly, for each possible multitemporal combination of modes, a prediction function $\widetilde{f}[\cdot]$ needs to be defined.

The contextual prediction function $\widetilde{f}[\cdot]$ can be expressed as a linear or nonlinear combination of the components of the vector of multitemporal observations x. While the latter ensures a more accurate prediction process, the former is usually preferred for its simplicity. In our case, another reason for adopting the linear prediction model is the fact that the prediction problem has been decomposed into easier prediction tasks for which a simple linear model could be sufficiently accurate.

Under a linear prediction model, the function $\widetilde{f}[\cdot]$ associated with the multitemporal combination of modes $(\omega_{n1}^1, \omega_{n2}^2, \ldots, \omega_{nK}^K)$ can be written as:

$$y = \tilde{f}[x_1, x_2, \ldots, x_K | x_1 \in \omega_{n1}^1, x_2 \in \omega_{n2}^2, \ldots, x_K \in \omega_{nK}^K] = \sum_{j=1}^{K} \beta_j \cdot x_j \quad (9.7)$$

where β_j stands for the weight assigned to the jth mode of the combination. In linear prediction, the determination of weight values represents the sole problem to face. This issue can be addressed in different ways. A simple solution widely used in the literature is based on the minimum square error pseudo-inverse technique [72]. This consists of solving the following system of R linear equations with K unknown variables ($R > K$):

$$\begin{bmatrix} x_1^1 & x_2^1 & \ldots & x_K^1 \\ x_1^2 & x_2^2 & \ldots & x_K^2 \\ \ldots & \ldots & \ldots & \ldots \\ x_1^R & x_2^R & \ldots & x_K^R \end{bmatrix} \cdot \begin{bmatrix} \beta_1 \\ \beta_2 \\ \ldots \\ \beta_K \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \ldots \\ y_R \end{bmatrix} \Leftrightarrow \mathbf{P} \cdot \beta = \mathbf{Y} \quad (9.8)$$

where R represents the number of multitemporal vectors $x^r = [x_1^r, x_2^r, \ldots, x_K^r]$ ($r \in \{1, 2, \ldots, R\}$) observed in $\overline{C}$ and collected in $\mathbf{P}$, such that $x_1^r \in \omega_{n1}^1, x_2^r \in \omega_{n2}^2, \ldots,$ and $x_K^r \in \omega_{nK}^K$. The corresponding observations y_r ($r \in \{1, 2, \ldots, R\}$) in the cloudy image $X^{(i)}$ (to be reconstructed) are gathered in the target vector $\mathbf{Y}$. The estimate of the optimal weight vector $\mathbf{fi}^*$ is given by the following equation based on the pseudoinverse $\mathbf{P}^\#$ of the matrix $\mathbf{P}$:

$$\beta = (\mathbf{P}^t \cdot \mathbf{P})^{-1} \cdot \mathbf{P}^t \cdot \mathbf{Y} = \mathbf{P}^\# \cdot \mathbf{Y} \tag{9.9}$$

The prediction system involved in the contextual multiple linear prediction (CMLP) method is thus made up of an ensemble of linear predictors, each trained to learn the relationship between images $X^{(k)}$ ($k \in S_c$) and image $X^{(i)}$ over a possible multitemporal combination of classes $(\omega_{n1}^1, \omega_{n2}^2, \ldots, \omega_{nK}^K)$ in $N(C)$. In addition, we integrate the ensemble with an additional linear predictor, termed global predictor, which is trained over all samples of $\overline{C}$ independently of their class membership. The motivation behind such an integration is that the global predictor is useful to deal with one of the following two possible situations: (1) a combination of classes in C does not exist in the set of feasible combinations of classes identified in $\overline{C}$; or (2) the number of samples collected in $\mathbf{P}$ and $\mathbf{Y}$ for a given multitemporal combination of classes available in $\overline{C}$ is not enough to apply in Eq. 9.9.

9.3.3.2 Unsupervised Classification with the EM Algorithm

EM Algorithm

As mentioned in the above algorithm, the first step of the contextual prediction process is that of classifying the region $N(C)$ of each image $X^{(k)}$ ($k \in S_c$) into a set of M_k data classes. This can be done in an unsupervised way (i.e., without the need of training samples) by means of any multilevel thresholding algorithm available in the literature, such as the basic minimum-error thresholding and the histogram peak selection algorithms [73, 74], or by using more sophisticated thresholding algorithms [75, 76].

> However, for histograms characterized by Gaussian mixtures, a particularly effective algorithm to detect the different modes accurately is the Expectation–Maximization (EM) algorithm [69, 70]. The EM algorithm is an iterative procedure that converges to local but usually good Maximum Likelihood (ML) estimates of mixture parameters.

Its convergence behavior is particularly well-studied for the classical case of Gaussian mixtures [77]. It is based on the interpretation of $\tilde{X}^{(k)} = \{x(u, v) : x(u, v) \in X^{(k)}, k \in S_c, (u, v) \in N(C)\}$ as incomplete data where the missing part is $Z^{(k)}$, i.e., its classification map. Assuming that L is the number of pixels in $\tilde{X}^{(k)}$, the missing

part can be evaluated as a set of L labels $Z^{(k)} = \{z_k^{(1)}, z_k^{(2)}, \ldots, z_k^{(L)}\}$ associated with the L pixels, indicating which class is at the origin of each pixel realization. Each label is a binary vector $z_k^{(i)} = [z_{k,1}^{(i)}, z_{k,2}^{(i)}, \ldots, z_{k,Mk}^{(i)}]$, such that $z_{k,r}^{(i)} = 1$ ($r \in 1, 2, \ldots, M_k$) if the ith pixel x_k^i of $\tilde{X}^{(k)}$ belongs to the rth data class ω_r^k, and $z_{k,r}^{(i)} = 0$ otherwise. The complete log-likelihood function, from which it would be possible to estimate the vector of parameters $\Theta^k = [P_1^k, P_2^k, \ldots, P_{M_k}^k, \mu_1^k, \mu_2^k, \ldots, \mu_{M_k}^k, \sigma_1^k, \sigma_2^k, \ldots, \sigma_{M_k}^k]$ if the complete data $\Psi^{(k)} = \{\tilde{X}^{(k)}, Z^{(k)}\}$ were observed, is given by:

$$\log p(\Psi^{(k)}|\Theta^k) = \ell(\Psi^{(k)}|\Theta^k) = \sum_{i=1}^{L}\sum_{r=1}^{M_k} z_{k,r}^{(i)} \log[P_r^k\, p(x_k^i|\theta_r^k)] \tag{9.10}$$

where $\theta_r^k = [\mu_r^k, \sigma_r^k]$.

The quantity $z_{k,r}^{(i)}$ can be estimated as the conditional expectation of $z_{k,r}^{(i)}$ given the observation $z_{k,r}^{(i)}$ and the set of parameters Θ^k [77]. The EM algorithm consists of expectation and maximization steps, which are iterated up to convergence. The expectation step is represented by the computations of $z_{k,r}^{(i)}$ ($i = 1, 2, \ldots, L$ and $r = 1, 2, \ldots, M_k$) using the current estimates of the set of parameters Θ^k. The maximization step allows updating such parameter estimates. It is possible to show that the equations related to these steps are as follows [77]:

1. E-step: compute $z_{k,r}^{(i)}$ given the parameter estimates from the previous M-step:

$$z_{k,r}^{(i)} = \frac{P_r^k \cdot N(x_k^i|\mu_r^k, \sigma_r^k)}{\sum_{j=1}^{M_k} P_j^k \cdot N(x_k^i|\mu_j^k, \sigma_j^k)} \tag{9.11}$$

2. M-Step: obtain new parameter estimates (denoted by the prime):

$$P'^k_r = \frac{1}{L}\sum_{i=1}^{L} z_{k,r}^{(i)} \tag{9.12}$$

$$\mu'^k_r = \frac{\sum_{i=1}^{L} z_{k,r}^{(i)} x_k^i}{\sum_{j=1}^{L} z_{k,r}^{(j)}} \tag{9.13}$$

$$\sigma'^k_r = \sqrt{\frac{\sum_{i=1}^{L} z_{k,r}^{(i)} (x_k^i - \mu_r'^k)^2}{\sum_{j=1}^{L} z_{k,r}^{(j)}}} \tag{9.14}$$

The values of the vector of parameters Θ^k used to start the EM algorithm can be chosen in different heuristic ways. In this work, we adopted a simple procedure that consists in dividing the histogram of $\tilde{X}^{(k)}$ into M_k regions of equal width fixed to:

$$\Delta = \frac{\underset{i=1,\ldots,L}{Max}\{x_k^i\} - \underset{i=1,\ldots,L}{Min}\{x_k^i\}}{M_k} \tag{9.15}$$

Then, the samples comprised in the rth histogram region bounded by the values $x_{left}^r = \underset{i=1,\ldots,L}{Min}\{x_k^i\} + (r-1)\cdot\Delta$ and $x_{right}^r = x_{left}^r + \Delta$ are used to compute the initial values of the three parameters $[P_r^k, \mu_r^k, \sigma_r^k]$ associated with the rth class ($r = 1, 2, \ldots, M_k$).

At convergence of the EM algorithm, the final parameter estimates will define completely the Gaussian data classes (modes) available in $\tilde{X}^{(k)}$. The latter is then transformed into a classification map with minimum error by adopting the maximum a posteriori probability (MAP) decision rule. Since the final estimates of $z_{k,r}^{(i)}$ represent the estimates of the posterior probabilities $P(\omega_r^k|x_k^i)$ ($i = 1, 2, \ldots, L$ and $r = 1, 2, \ldots, M_k$), one can assign to each pixel x_k^i of $\tilde{X}^{(k)}$ the optimal class label $\hat{\omega} \in \Omega = \{\omega_r^k : r = 1, 2, \ldots, M_k\}$, such that:

$$\hat{\omega} = \underset{\omega_r^k \in \Omega}{argmax}\, P(\omega_r^k|x_k^i) \tag{9.16}$$

Estimation of the Number of Classes

As the number of data classes M_k is not known a priori and therefore needs to be estimated, we have to resort to a technique that deals with this important issue, which is typical of mixture modeling problems.

> Indeed, the selection of the number of components in a mixture raises a tricky trade-off, since on the one hand the higher the number of components, the higher the risk of data overfitting, while on the other, the smaller the number of components, the lower the model flexibility. In the literature, the most popular methods for estimating automatically the number of data classes are based on approximate Bayesian criteria or on information theory concepts [77].

In this work, we will use the Minimum Description Length (MDL) criterion, which takes origin from the information theory and is defined as [78]:

$$MDL(M_k) = -\tilde{\ell}(\Psi^{(k)}|\Theta^k) + \gamma\cdot\kappa\cdot\log(L) \tag{9.17}$$

where $\tilde{\ell}(\Psi^{(k)}|\Theta^k)$ represents the log-likelihood function value found at convergence of the EM algorithm, κ is the number of parameters in Θ^k, and γ is a constant. In our case, since there are three parameters to estimate (prior, mean, and standard deviation)

to define the Gaussian distribution associated with each data class completely, κ is given by:

$$\kappa = 3 \cdot M_k - 1 \tag{9.18}$$

The "-1" term in Eq. 9.18 is explained by the fact that the constraint $\sum_{i=1}^{M_k} P_i^k = 1$ allows saving a free parameter. For the setting of γ, different values are used in the literature. According to [79], $\gamma = 5/2$ seems the most appropriate choice.

The optimal number of data classes $\hat{M}_k$ in $\tilde{X}^{(k)}$ is estimated by minimizing the MDL criterion, i.e.,

$$\hat{M}_k = \underset{M_k=1,\ldots,M_{\max}}{argmin} \{MDL(M_k)\} \tag{9.19}$$

where M_{max} is a predefined maximal number of data classes.

9.3.4 Compressive Sensing Reconstruction Strategies

The second reconstruction method is based on compressive sensing (CS). CS relies on the idea to exploit redundancy (if any) in the signals and represent them in a sparse way [80, 81]. Usually, signals like images are sparse, as they contain, in some representation domain, many coefficients close to or equal to zero. CS starts taking a weighted linear combination of pixels in a basis in which the signal is assumed to be sparse.

The fundamental of the CS theory is the ability to recover with relatively few measurements $x = D \cdot \alpha$ by solving the following ℓ_0-minimization problem:

$$\min ||\alpha||_0 \text{ subject to } x = D \cdot \alpha \tag{9.20}$$

where D is a dictionary with a certain number of atoms, x is the original signal which can be represented as a sparse linear combination of these atoms, and the minimization of $|| \cdot ||_0$, the ℓ_0-norm, corresponds to the maximization of the number zeros in α, following this formulation: $||\alpha||_0 = \#\{i : \alpha_i \neq 0\}$.

Equation 9.20 represents a NP-hard problem, which means that it is computationally infeasible to solve. Following the discussion of Cands and Tao [82], it is possible to simplify the evaluation of Eq. 9.20 in a relatively easily linear programming solution. They demonstrate that, under some reasonable assumptions, minimizing ℓ_1-norm is equivalent to minimizing ℓ_0-norm, which is defined as $||\alpha||_1 = \sum_i |\alpha_i|$. Accordingly, it is possible to rewrite Eq. 9.20 as:

$$\min ||\alpha||_1 \text{ subject to } x = D \cdot \alpha \tag{9.21}$$

In the literature, there exist several algorithms for solving optimization problems similar to the one expressed in Eq. 9.21. In the next subsection, we briefly introduce two of them, which represent the most common solutions from the literature.

9.3.4.1 CS Solutions

A well-known solution for problem expressed in Eq. 9.21 is the basis pursuit (BP) principle [82, 83]. It suggests a convexification of the problem by using the ℓ_1-norm instead of ℓ_0.

This means that the best approximation of the problem becomes equal to a support minimization problem. BP finds signal representations in overcomplete dictionaries by convex, nonquadratic optimization technique, solving problem in Eq. 9.21. It can be reformulated as a linear programming (LP) problem, and solved using modern interior-point methods, simplex methods, or other techniques, such as homotopy techniques [84]. Given that, it is possible to rewrite the ℓ_1-norm in Eq. 9.21 as:

$$||\alpha||_1 = \sum_i |\alpha_i| = \sum_i \mu_i + v_i \text{ where } \begin{cases} \alpha_i = u_i, v_i = 0 \text{ if } \alpha_i \geq 0 \\ \alpha_i = -v_i, u_i = 0 \text{ if } \alpha_i \leq 0 \end{cases} \tag{9.22}$$

Substituting it in Eq. 9.21, it allows to perform a linear minimization problem. Note that, if the original signal x is sufficiently sparse, the recovery via BP is probably exact.

One of the easiest and fastest alternative technique is the orthogonal matching pursuit (OMP), an improved version of the matching pursuit (MP) method. MP finds the atom that has the highest correlation with the signal. It subtracts off the correlated part from the signal and then iterates the procedure on the resulting residual signal [85, 86].

The algorithm approximates the signal x, considering these two decompositions [83]:

$$x = \sum_{d \in D} \alpha_d \phi_d = \sum_{i=1}^{m} \alpha_{d_i} \phi_{d_i} + R^{(m)} \tag{9.23}$$

where the dictionary D is a collection of atom vectors $\{\phi_d\}_{d \in D}$ and $R^{(m)}$ is a residual. Starting from an initial approximation $x^{(0)} = 0$ and residual $R^{(0)} = x$, it builds up a sequence of sparse approximations stepwise. At stage k, it identifies the dictionary

atom that best correlates with the residual and then adds to the current approximation, a scalar multiple of that atom, so that $x^{(k)} = x^{(k-1)} + \alpha_k \phi_{d_k}$, where $\alpha_k = \left\langle R^{(k-1)}, \phi_{d_k} \right\rangle$ and $R^{(k)} = x - x^{(k)}$. After m steps, one has a representation of the form of Eq. 9.23, with residual $R = R^{(m)}$, where the original signal x is decomposed into a sum of dictionary elements that are chosen to best match its residues. Unfortunately, the convergence speed of this algorithm is not fast. To overcome this drawback, an improved solution called orthogonal MP (OMP) was developed. Differently from MP, OMP updates the coefficients of the selected atoms at each iteration so that the resulting residual vectors are orthogonal to the subspace spanned by the selected atoms. When stopped after only few iterations, it generally yields a satisfactory approximation, using only few atoms [85, 86].

From the literature [87, 88], it comes out that BP and OMP algorithms provide in general good performances in reconstruction problems. Nonetheless, BP is considered more powerful than OMP, since it can recover with high probability all sparse signals and is more stable. On the contrary, OMP results attractive for its fast convergence and in its ease of implementation.

9.3.4.2 Genetic Algorithm-Based CS Solution

A third CS strategy that we investigate in this part of the chapter is based on genetic algorithms (GAs).

> GAs are a part of evolutionary computation which solves optimization problem by mimicking the principles of biological evolution [89, 90]. A genetic optimization algorithm performs a search by regenerating a population of candidate solutions (or individuals) represented by chromosomes. From one generation to the next, the population is improved following biological rules, adopting deterministic, and nondeterministic genetic operators.

In general, a common GA involves the following steps. First, an initial population of chromosomes is randomly generated. Then, the goodness of each chromosome is evaluated according to a predefined fitness function representing the aim of the optimization. Evaluating the fitness function allows to keep or discard chromosomes by using a proper rule based on the principle that the better the fitness, the higher the chance of being selected. Once the selection of the best chromosomes is done, the next step is devoted to the reproduction of a new population. This is done by genetic operators such as crossover and mutation operators. All these steps are iterated until some predefined condition is satisfied (e.g., maximum number of generations or fitness value limit).

Several multiobjective GA-based approaches have been proposed in the literature [91], such as SPEA-II [92], PAES [93], and NSGA-II [94]. We will adopt the nondominated sorting solution (NSGA-II) for its low computational requirements, its

aptitude to distribute uniformly the optimal solutions along the Pareto front [94], and its successful application to different remote sensing problems [95–97]. NSGA-II is based on the concept of nondominance. A solution s_1 is said to dominate another solution s_2 if s_1 is not worse than s_2 in all objectives and better than s_2 in at least one objective. A solution is said to be nondominated if it is not dominated by any other solution. As GA does, NSGA-II starts by generating a random parent population. Individuals (chromosomes) selected through a crowded tournament selection undergo crossover and mutation operations to form an offspring population. Both offspring and parent population are then combined and sorted into fronts of decreasing dominance (rank). After the sorting process, the new population is filled with solutions of different fronts starting from the best one. If a front can only partially fill the next generation, crowded tournament selection is applied again to ensure diversity. Once the next generation population has been completely filled, the algorithm loops back to create a new offspring population and the process continues up to convergence.

The design of a multiobjective GA optimization relies upon two components, the chromosome structure and the fitness functions, which encode the considered optimization problem and show the direction to obtain the best solution, respectively.

Concerning the first component, we consider a population of M chromosomes $\alpha_m \in \Re,\ m \in \{1, \ldots, M\}$, where each chromosome is a real vector composed of genes corresponding to the weight vector α defined above in the previous sections. The length w of the chromosome is thus equal to the one of the dictionary D. Chromosomes can be randomly initialized or, to obtain a faster and better convergence, it could be envisioned to add a priori information coming from more simple CS techniques, i.e., OMP and BP algorithms. In other words, one could exploit OMP and BP solutions to generate an initial population by perturbing these solutions.

Regarding the fitness function, we investigate separately and jointly two fitness functions, i.e., those defining the optimization problem in Eq. 9.20. The first one aims to maximize the sparsity level by minimizing the ℓ_0-norm of the weight vector α, which corresponds to minimize the number of almost nonzero coefficients in α:

$$f_1 = \min ||\alpha||_0 \tag{9.24}$$

An almost nonzero coefficient is a coefficient exhibiting a value less than a predefined small threshold value (th). The second fitness function is derived from the constraint in Eq. 9.20. It points to a perfect reconstruction of the considered pixel (at position $[k, l]$). In other words, it is expressed using the 2-norm as:

$$f_2 = \min ||D\alpha - x||^2. \tag{9.25}$$

NSGA-II returns several optimal (nondominated) solutions along the Pareto front. Since a single solution has to be selected from the nondominated set, different strategies can be adopted. In this study, we suggest to choose the median solution as typically performed in the literature. In such a way, we expect to get a compromise solution with respect to what could be obtained by OMP and BP, i.e., a trade-off between reconstruction model sparseness and reconstruction error.

9.3.5 Illustration Examples

9.3.5.1 Data set Description and Experimental Design

The two investigated reconstruction methods have been evaluated experimentally on real multitemporal multispectral remote sensing images. The first data set comes from the Taiwanese optical high resolution FORMOSAT-2 satellite, which permits to acquire repeat imagery of an area of interest every day, from the same angle and under the same light conditions [98]. These images represent part of the Arcachon basin in the south region of Aquitaine, in France. The images are composed of 400×400 pixels, 4 spectral bands (blue, green, red and near-infrared) with a spatial resolution of 8 meters. They were acquired on the June 24 and the July 16, 2009, respectively (see cropped version in Fig. 9.6). The second data set comes from the SPOT-5 French satellite, whose images represent part of the Reunion Island [99]. The images are characterized by a size of 450×450 pixels, 4 spectral bands (blue, green, red and near-infrared), a spatial resolution of 10 meters and were taken on the 2nd of May and the 18th of June 2008, respectively (data not shown). The two data sets present

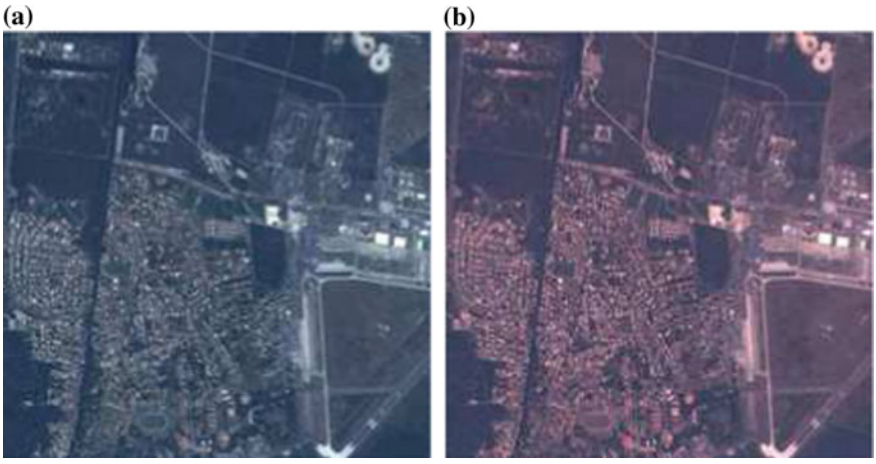


Fig. 9.6 Data set 1: Crop of the FORMOSAT-2 images acquired in the Arcachon area on **a** the 24th of June and **b** the 16th of July, 2009

several differences such as sensor type, spatial resolution, and types of land covers. Indeed, the first one is dominated by vegetation areas, while the second one presents more urban areas. In terms of comparisons, we considered also a recent work based on multiresolution inpainting (MRI) [100].

In order to quantify the reconstruction accuracy, the experiments were done in this way: (1) a cloud-free image $I^{(1)}$ is considered; (2) the presence of clouds is simulated by partly obscuring the other image $I^{(2)}$; and (3) the reconstructed image is compared with the original cloud-free image. The simulation study is aimed at understanding the sensitivity of the five investigated methods (i.e., the MRI, the CMLP, the OMP, the BP, and the GA) to two aspects: (1) the kind of ground covers obscured and (2) the size of the contaminated area. In order to obtain a detailed assessment of the reconstruction quality, we adopt the well-known peak signal-to-noise ratio (PSNR) measure [100]. Another quantitative criteria is the computational time (in seconds). Regarding the dictionaries, we collected directly training samples from the image, by sampling pixels in the source region Φ. For the GA setup, we used the following parameters: chromosome number $M = 50$, threshold value $th = 10^{-4}$, probability of crossover $P_c = 0.8$, and probability of mutation $P_m = 0.005$.

9.3.5.2 Simulation Experimental Results

Contamination of Different Ground Cover

Different masks with different positions were considered in a way to simulate the obscuration of different kinds of ground cover. In particular, for data set 1, mask A covered a region that included mainly a urban area, mask B obscured an industrial zone, and mask C covered a vegetation area. For data set 2, mask A covered mainly a rural area, and mask B a vegetation region. The experiments were carried out by considering each mask at a time, where each mask was composed by around 2000 pixels and the dictionary by 300 pixels.

Table 9.1 reports the results of the different reconstruction techniques over different obscured land covers. In greater detail, MRI generally reconstructs the missing data with a good PSNR level, but the corresponding reconstructed images appear visually of poor quality since it does not capture satisfactorily the textural properties of the missing areas. In general, MRI can return visually satisfactory results only when the missing area refers to a uniform region such as vegetation region. This is the case for mask C in data set 1 and mask B in data set 2. CMLP method provides generally satisfactory results in terms of reconstruction error and computation time. To obtain better results, it would need more than two temporal images. Coming now to the CS-based implementations, OMP algorithm produces very sparse reconstruction solutions (around 3 nonzero coefficients). On the one hand, this may be an advantage in terms of computation time. On the other hand, OMP is potentially subject to underfitting problems. On the contrary, BP algorithm may be subject to overfitting problems due to the fact that most of the time it selects a large number of weight coefficients (in general around 300 coefficients). Comparing OMP and

Table 9.1 Quantitative results obtained in the first simulation experiments for (a) the first and (b) the second data set

(a)

	Mask A			Mask B			Mask C		
Method	PSNR		Time [s]	PSNR		Time [s]	PSNR		Time [s]
	l_1	l_2		l_1	l_2		l_1	l_2	
BP	80.59	22.22	66	77.10	24.74	59	98.53	30.67	60
CMLP	-	20.99	1	-	20.11	1	-	24.05	1
GA	42.09	23.78	68621	43.38	23.15	26312	45.62	32.01	43193
MRI	-	22.54	2856	-	16.05	2517	-	33.77	2898
OMP	39.41	23.96	4	36.33	20.60	4	44.28	31.97	4

(b)

	Mask A			Mask B			
Method	PSNR		Time [s]	PSNR		Time [s]	
	l_1	l_2		l_1	l_2		
BP	86.22	26.45	61	99.62	31.63	91	
CMLP	-	24.61	1	-	27.69	1	
GA	50.70	26.72	69231	56.30	31.28	38475	
MRI	-	24.27	2995	-	29.54	3614	
OMP	46.53	26.36	5	54.49	30.43	5	

BP in terms of computation time, the latter is far less efficient, whereas in terms of PSNR, both methods return similar reconstruction values, outperforming CMLP and MRI. Lastly, GA can be viewed as a compromise between the two previous methods. Despite the very long time needed to estimate the reconstruction model, it results sparser than BP, but less parsimonious to OMP. Its reconstruction error is almost all the time the best or the second best.

Contamination with Different Sizes

Another important test for the five methods consists of assessing their performances by varying the amount of missing data. Three different masks were adopted to simulate increasing cloud cover sizes. In particular, mask 1 was the same as the mask A adopted in the previous experiments, i.e., it covered about 2000 pixels. To build masks 2 and 3, we multiplied the previous size by 3 and by 6. Also in these experiments, the adopted dictionaries were composed of 300 pixels belonging to the Φ region. Table 9.2 reports the results achieved by the different reconstruction techniques and by varying the amount of missing data.

From a quantitative viewpoint, in terms of PSNR, we have similar results as in the previous experiments. MRI still presents problems in reconstructing satisfactorily complex textures. CMLP competes seriously with MRI in terms of computation time and PSNR. However to get higher PSNR values, one needs to resort to CS techniques. Indeed, the CS-based implementations return better results in term of

Table 9.2 Quantitative results obtained in the second simulation experiments for (a) the first and (b) the second data set

(a)

	Mask A			Mask B			Mask C		
Method	PSNR		Time [s]	PSNR		Time [s]	PSNR		Time [s]
	l_1	l_2		l_1	l_2		l_1	l_2	
BP	80.59	22.22	66	80.18	22.89	145	79.53	21.47	865
CMLP	-	20.99	1	-	21.13	1	-	20.83	2
GA	42.09	23.78	68621	45.46	23.85	99072	45.13	23.03	275394
MRI	-	22.54	2856	-	21.35	6938	-	19.63	14774
OMP	39.41	23.96	4	42.45	23.21	6	42.00	25.01	19

(b)

	Mask A			Mask B			Mask C		
Method	PSNR		Time [s]	PSNR		Time [s]	PSNR		Time [s]
	l_1	l_2		l_1	l_2		l_1	l_2	
BP	86.22	26.45	61	87.49	26.82	143	86.60	28.25	972
CMLP	-	24.61	1	-	24.43	2	-	25.46	2
GA	50.70	26.72	69231	50.63	27.10	103342	51.14	28.15	259459
MRI	-	24.27	2995	-	22.85	10176	-	23.82	22353
OMP	46.53	26.36	5	46.89	26.42	16	47.49	27.39	21

PSNR in all the simulations and present the advantage for not depending on the size of the missing area. The best solution in these experiments in terms of PSNR comes from GA, which outperforms all other methods in three cases, and in the other three, it is the second best choice. About the computation time, as expected, it increases as the amount of missing data increases. Results from this viewpoint underline the main weakness of the GA solution, i.e., its expensive computational needs.

Figure 9.7 shows qualitative reconstruction results in RGB composites obtained in the most critical reconstruction scenario, i.e., the largest simulated cloud mask 3, for the first data set for all reconstruction methods. As mentioned before, MRI reconstruction exhibits the worst reconstruction case. CMLP method is capable to obtain a good reconstruction compared with MRI. Regarding the CS reconstruction techniques (OMP, BP, and GA), good reconstructions are obtained where it is not simple to find significant differences comparing the reconstructions with the original (cloud-free) image.

Reconstruction Impact on Image Classification

Since the generation of classification maps represents one of the most widespread applications of remote sensing images, it was also worth to evaluate the quality of the reconstruction process in terms of classification error. The latter was computed first by generating a classification map of the original images that served as reference classification maps by means of the popular *k*-means classifier. Then, each reconstructed image was given in input to the k-means classifier to provide a reconstruction classi-

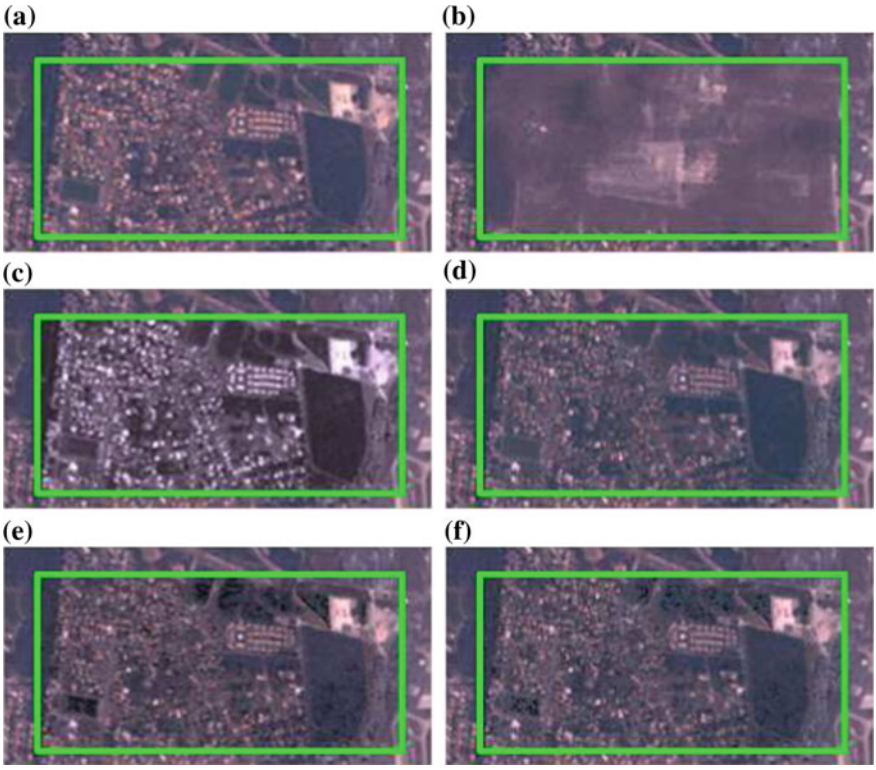


Fig. 9.7 Data set 1. Color composite images (bands 1, 2, and 3) **a** of the original image and the same image reconstructed after the contamination with the largest simulated mask 3 by **b** MRI, **c** CMLP, **d** OMP, **e** BP, and **f** GA methods (colour figure online)

fication map. For each reconstruction method, it was possible to evaluate the overall number of classification errors (OE) inside the reconstructed cloud-contaminated area by a simple comparison of both the reconstruction and the original classification maps. We repeated this exercise with different numbers of clusters (from 3 to 7 clusters). The results confirm what previously observed, i.e., CS methods behave better than the other methods. As example, we have reported in Fig. 9.8 the clustering results (with $k = 5$) obtained for the FORMOSAT-2 image with mask A. The best classification is achieved from the reconstruction with OMP (OE of 6.3%), followed by GA (OE = 8.7%), BP (OE = 19.6%), MRI (OE = 25.2%), and CMLP (OE = 29.4%).

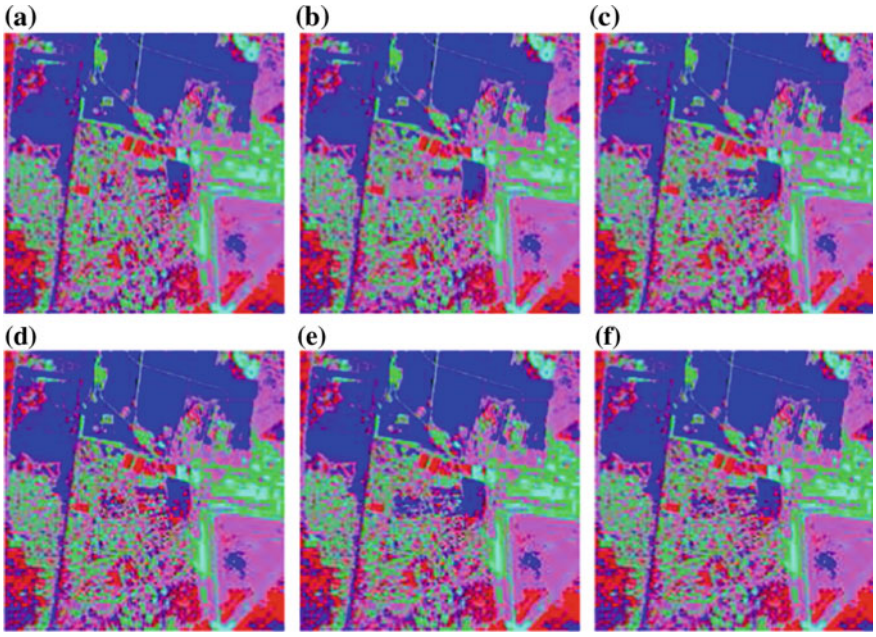


Fig. 9.8 Unsupervised classification maps obtained by the k-means algorithm ($k = 5$) **a** for the original FORMOSAT-2 image and the same image reconstructed after contamination with mask A by **b** MRI, **c** CMLP, **d** OMP, **e** BP, and **f** GA methods

9.3.6 Discussion

In this part of the chapter, we have dealt with the complex and important problem of the removal of clouds from sequences of multitemporal multispectral optical images. Given a contaminated image of a sequence, each area of missing measurements is reconstructed. Two main approaches have been investigated.

The first method is based on a contextual prediction system, which is trained in an unsupervised way to reproduce the local spectro-temporal relationships between the considered image and an opportunely selected subset of the remaining temporal images. The prediction system characterizing the CMLP method is based on an ensemble of contextual linear predictors, each associated with a local temporal region spectrally homogeneous in each image of the selected subset. These regions are identified by an EM-based unsupervised classifier. The main properties of the method are the following: (1) it relies on the assumptions that spectral non-stationarity is allowed, while the spatial structure of the image should be almost identical over the image sequence; (2) it is not ground cover-dependent; (3) it is not based on the compositing principle, i.e., it allows the reconstruction of each image of the sequence separately; (4) it is completely unsupervised; (5) it is conceptually simple, easy to implement, and relatively fast to run; (6) it achieves a satisfactory reconstruction

quality with regard to the complexity of the faced ill-posed problem; (7) it can be used to recover measurements that may be due not only to clouds but also to shadows.

The experimental results point out the kind of ground cover obscured and the size of the contaminated area only marginally affect the performance of the reconstruction method. The latter depends more directly on the representativeness of the samples extracted from the spatial neighborhood of the contaminated area and used to train their predictor(s). In other words, if a ground cover is contaminated and it is not represented in the neighborhood of the contamination area, the contextual prediction process will not be capable of dealing suitably with such a situation. The assessment of the method in terms of classification accuracy, considered by the remote sensing community as a fundamental criterion for quality evaluation, confirms that the reconstruction process is capable of capturing with satisfactory accuracy the trends of the true statistical model associated with missing data. Furthermore, results show that the impact of the recovered samples on the statistical structure of the remaining part of the image is very limited, thus underlining the high statistical compatibility between the recovered samples and the uncontaminated ones.

In order to improve the accuracy of the reconstruction process, different aspects of the method deserve to be investigated including, for instance, the problem of the independent reconstruction of single-channel images. In this context, the second approach, which is based on compressive sensing (CS), overcomes this issue through the implementation of a joint multichannel reconstruction process. In particular, three different strategies have been investigated. First, we have shown how two common CS solutions, the orthogonal matching pursuit (OMP) and the basis pursuit (BP) algorithms, can be formulated for a cloud-contaminated image reconstruction problem. Then, we have seen a solution for solving the CS problem under a L_0-norm perspective, exploiting the capabilities of genetic algorithms (GAs). The main properties of the methods are: (1) they rely on the assumptions that spectral non-stationarity is allowed, while the spatial structure of the image should be almost identical between the two images; (2) they are not ground cover-dependent; (3) they are unsupervised; (4) differently from CMLP, they need just one reference cloud-free image and, as mentioned above, the reconstruction of each pixel is performed in all spectral bands simultaneously.

9.4 Conclusion

In this chapter, we have described two main issues related to the processing of Satellite Image Time Series and proposed different approaches to deal with them. The first part of the chapter is focused on data mining methods for extracting spatiotemporal patterns in an unsupervised way. A data mining technique which can handle various kind of SITS is illustrated on time series of displacement measurements derived from multitemporal InSAR image. It requires only three parameters and has been shown to be useful to extract phenomena that could not be unveiled by other approaches. In particular, GFS-patterns can refine one another spatially and temporally: they can

overlap one another both in time and in space. Segmentation or clustering techniques cannot give access to such descriptions. Another interesting property is the ability of the method to discard random uncertainty such as atmospheric turbulences. Numerous experiments such as those reported in [31, 32, 34, 38, 40, 41] confirm these properties. Besides human interpretation, the proposed technique can also be used for SITS indexing and retrieval.

In the second part of the chapter, we have investigated two approaches to deal with the removal of clouds from sequences of multitemporal multispectral optical images. The first strategy is based on a contextual prediction system, while the second is based on CS. The experimental results point out the superiority of the CS approach. Comparing the CS solutions, OMP has the advantage to be sparser and significantly faster than BP and GA, but it is the less robust method. Indeed, since the reconstruction of each pixel depends typically on 3 coefficients and thus 3 other pixels of the image, it is enough that one of them is missing (covered by a cloud) to render the reconstruction model inaccurate. This problem is much less important to BP as it is much less sparse than OMP. GA represents a good compromise between OMP and BP methods, mainly because it is more robust than OMP and more sparse than BP. Another empirical conclusion is that the kind of ground cover obscured may be an important factor to take in consideration for the reconstruction, while the size of the contaminated area only marginally affects the performance of the explored reconstruction methods, which depend more on the information available outside the missing area. In other words, if a ground cover is contaminated and it is not represented outside of the contaminated area, the reconstruction process will not deal with such a situation correctly.

References

1. Lopès, A., Garello, R., Le Hégarat-Mascle, S.: Speckle models. Processing of Synthetic Aperture Radar (SAR) Images. Wiley-ISTE, New York (2008)
2. Trouvé, E., Chambenoit, Y., Classeau, N., Bolon, P.: Statistical and operational performance assessment of multitemporal SAR image filtering. IEEE Trans. Geosci. Remote Sens. **41**(11), 2519–2530 (2003)
3. Ciuc, M., Bolon, P., Trouvé, E., Buzuloiu, V., Rudant, J.-P.: Adaptive-neighborhood speckle removal in multitemporal synthetic aperture radar images. Appl. Opt. **40**(32), 5954–5966 (2001)
4. Bruniquel, J., Lopès, A.: Multi-variate optimal speckle reduction in SAR imagery. Int. J. Remote Sens. **18**(3), 603–627 (1997)
5. Atto, A.M., Trouvé, E., Nicolas, J.-M., Le, T.T.: Wavelet operators and multiplicative observation models -application to SAR image time series analysis. IEEE Trans. Geosci. Remote Sens. (2016). https://hal.archives-ouvertes.fr/hal-01341064
6. Le, T.T., Atto, A.M., Trouvé, E., Nicolas, J.-M.: Adaptive multitemporal SAR image filtering based on the change detection matrix. IEEE Geosci. Remote Sens. Lett. **11**(10), 5 (2014). https://hal.archives-ouvertes.fr/hal-00975385
7. Su, X., Deledalle, C.A., Tupin, F., Sun, H.: Two-step multitemporal nonlocal means for synthetic aperture radar images. IEEE Trans. Geosci. Remote Sens. **52**(10), 6181–6196 (2014)

8. Bujor, F., Trouvé, E., Valet, L., Nicolas, J.-M., Rudant, J.-P.: Application of log-cumulants to the detection of spatiotemporal discontinuities in multitemporal SAR images. IEEE Trans. Geosci. Remote Sens. **42**(10), 2073–2084 (2004)
9. Ferretti, A., Prati, C., Rocca, F.: Permanent scatterers in SAR interferometry. IEEE Trans. Geosci. Remote Sens. **39**(1), 8–20 (2001)
10. Berardino, P., Fornaro, G., Lanari, R., Sansosti, E.: A new algorithm for surface deformation monitoring based on small baseline differential SAR interferograms. IEEE Trans. Geosci. Remote Sens. **40**, 2375–2382 (2002)
11. Lu, D., Mausel, P., Brondizio, E., Moran, E.: Change detection techniques. Int. J. Remote Sens. **25**(12), 2365–2401 (2004)
12. Coppin, P., Jonckheere, I., Nackaerts, K., Muys, B., Lambin, E.: Review articledigital change detection methods in ecosystem monitoring: a review. Int. J. Remote Sens. **25**(9), 1565–1596 (2004)
13. Petitjean, F., Inglada, J., Gançarski, P.: Satellite image time series analysis under time warping. IEEE Trans. Geosci. Remote Sens. **50**(8), 3081–3095 (2012)
14. Millward, A.A., Piwowar, J.M., Howarth, P.J.: Time-series analysis of medium-resolution, multisensor satellite data for identifying landscape change. Photogramm. Eng. Remote Sens. **72**(6), 653–663 (2006)
15. Foody, G.: Monitoring the magnitude of land-cover change around the southern limits of the Sahara. Photogramm. Eng. Remote Sens. **67**(7), 841–848 (2001)
16. Melgani, F., Moser, G., Serpico, S.B.: Unsupervised change-detection methods for remote-sensing images. Opt. Eng. **41**(12), 3288–3297 (2002)
17. Inglada, J., Mercier, G.: A new statistical similarity measure for change detection in multitemporal SAR images and its extension to multiscale change analysis. IEEE Trans. Geosci. Remote Sens. **45**(5), 1432–1445 (2007)
18. Lambin, E.F., Strahlers, A.H.: Change-vector analysis in multitemporal space: a tool to detect and categorize land-cover change processes using high temporal-resolution satellite data. Remote Sens. Environ. **48**(2), 231–244 (1994)
19. JHA, C.S., Unni, N.: Digital change detection of forest conversion of a dry tropical indian forest region. Int. J. Remote Sens. **15**(13), 2543–2552 (1994)
20. Julea, A., Méger, N., Bolon, P., Rigotti, C., Doin, M.-P., Lasserre, C., Trouvé, E., Lăzărescu, V.N.: Unsupervised spatiotemporal mining of satellite image time series using grouped frequent sequential patterns. IEEE Trans. Geosci. Remote Sens. **49**(4), 1417–1430 (2011)
21. Celik, T., Ma, K.-K.: Multitemporal image change detection using undecimated discrete wavelet transform and active contours. IEEE Trans. Geosci. Remote Sens. **49**(2), 706–716 (2011)
22. Frawley, W., Piatetsky-Shapiro, G., Matheus, C.: Knowledge discovery in databases: an overview. In: Piatetsky-Shapiro, G., Frawley, W. (eds.) Knowledge in Discovery in Databases, pp. 1–27. AAAI Press, Menlo Park (1991)
23. Honda, R., Konishi, O.: Temporal rule discovery for time-series satellite images and integration with RDB. In: PKDD '01: Proceedings of the 5th European Conference on Principles of Data Mining and Knowledge Discovery, pp. 204–215. Springer, London (2001)
24. Héas, P., Datcu, M.: Modeling trajectory of dynamic clusters in image time-series for spatio-temporal reasoning. IEEE Trans. Geosci. Remote Sens. **43**(7), 1635–1647 (2005)
25. Nezry, E., Genovese, G., Solaas, G., Rémondière, S.: ERS - Based early estimation of crop areas in Europe during winter 1994–95. In: Guyenne, T.-D. (ed.) ERS Application, Proceedings of the Second International Workshop held 6–8 December 1995 in London, vol. 383, p. 13–20. ESA Special Publication (1996)
26. Petitjean, F., Inglada, J., Gancarski, P.: Satellite image time series analysis under time warping. IEEE Trans. Geosci. Remote Sens. **50**(8), 3081–3095 (2012)
27. Gueguen, L., Datcu, M.: Image time-series data mining based on the information-bottleneck principle. IEEE Trans. Geosci. Remote Sens. **45**(4), 827–838 (2007)
28. Gallucio, L., Michel, O., Comon, P.: Unsupervised clustering on multi-components datasets: applications on images and astrophysics data. In: 16th European Signal Processing Conference EUSIPCO-2008, Lausanne, Switzerland, August 2008, pp. 25–29 (2008)

29. Agrawal, R., Srikant, R.: Mining sequential patterns. In: Yu, P.S., Chen, A.S.P. (eds.) Proceedings of the 11th International Conference on Data Engineering (ICDE'95), pp. 3–14. IEEE Computer Society Press, Taipei, Taiwan (1995)
30. Luo, C., Chung, S.-M.: Efficient mining of maximal sequential patterns using multiple samples. In: Proceedings of the 2005 SIAM International Conference on Data Mining (2005)
31. Julea, A., Méger, N., Bolon, P., Rigotti, C., Doin, M.-P., Lasserre, C., Trouvé, E., Lăzărescu, V.: Unsupervised spatiotemporal mining of satellite image time series using grouped frequent sequential patterns. IEEE Trans. Geosci. Remote Sens. **49**(4), 1417–1430 (2011)
32. Julea, A., Méger, N., Rigotti, C., Trouvé, R., Jolivet, Emmanuel, Bolon, P.: Efficient spatiotemporal mining of satellite image time series for agricultural monitoring. Trans. Mach. Learn. Data Min. **5**(1), 23–44 (2012)
33. Meger, N., Jolivet, R., Lasserre, C., Trouve, E., Rigotti, C., Lodge, F., Doin, M.-P., Guillaso, S., Julea, A., Bolon, P.: Spatiotemporal mining of ENVISAT SAR interferogram time series over the Haiyuan fault in China. In: 2011 6th International Workshop on the Analysis of Multi-temporal Remote Sensing Images (Multi-Temp), pp. 137–140 (2011)
34. Rigotti, C., Lodge, F., Meger, N., Pothier, C., Jolivet, R., Lasserre, C.: Monitoring of tectonic deformation by mining satellite image time series. In: 19th National Conference Reconnaissance de Formes et Intelligence Artificielle (RFIA'14), July 2014, pp. 1–6 (2014)
35. Good, P.: Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses. Springer Series in Statistics. Springer, Berlin (2000)
36. Cobb, G.W., Chen, Y.-P.: An application of Markov chain Monte Carlo to community ecology. Am. Math. Monthly **110**(4), 265–288 (2003)
37. Gionis, A., Mannila, H., Mielikäinen, T., Tsaparas, P.: Assessing data mining results via swap randomization. ACM Trans. Knowl. Discov. Data **1**(3) (2007)
38. Méger, N., Rigotti, C., Pothier, C.: Swap randomization of bases of sequences for mining satellite image times series. Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2015, Porto, Portugal, 7–11 September 2015, Proceedings, Part II, pp. 190–205. Springer International Publishing, Berlin (2015)
39. Cover, T., Thomas, J.: Elements of Information Theory. Wiley, New York (1991)
40. Lodge, F., Meger, N., Rigotti, C., Pothier, C., Doin, M.-P.: Iterative summarization of satellite image time series. In: IEEE International Geoscience and Remote Sensing Symposium, July 2014, pp. 1–4 (2014)
41. Meger, N., Rigotti, C., Gueguen, L., Lodge, F., Pothier, C., Andreoli, R., Datcu, M.: Normalized mutual information-based ranking of spatio-temporal localization maps. In: 8th European Spatial Agency (ESA) - EUSC - JRC Conference on Image Information Mining, October 2012, pp. 11–14 (2012)
42. SITS-P2MINER: https://int.polytech.univ-smb.fr/fileadmin/polytech_autres_sites/sites/listic/projets/sitsmining/SITSP2MINER.zip (2016). Accessed 12 Apr 2016
43. Cihlar, J.: Remote sensing of global change: an opportunity for Canada. In: Proceedings of 11th Canadian Symposium on Remote Sensing, pp. 39–48 (1987)
44. Wang, J.R., Racette, P., Triesky, M., Browell, E., Ismail, S., Chang, L.A.: Profiling of atmospheric water vapor with MIR and LASE. IEEE Trans. Geosci. Remote Sens. **40**(6), 1211–1219 (2002)
45. Vasudevan, B.G., Gohil, B.S., Agarwal, V.K.: Backpropagation neural-network-based retrieval of atmospheric water vapor and cloud liquid water from IRS-P4 MSMR. IEEE Trans. Geosci. Remote Sens. **42**(5), 985–990 (2004)
46. Stowe, L.L., Davis, P.A., McClain, E.P.: Scientific basis and initial evaluation of the CLAVR-1 global clear/cloud classification algorithm for the advanced very high resolution radiometer. J. Atmos. Ocean. Technol. **16**(6), 656–681 (1999)
47. Sea, B., Channel, A.: The cloud and surface parameter retrieval (CASPR) system for polar AVHRR (2002)
48. Di Vittorio, A.V., Emery, W.J.: An automated, dynamic threshold cloud-masking algorithm for daytime AVHRR images over land. IEEE Trans. Geosci. Remote Sens. **40**(8), 1682–1694 (2002)

49. Logar, A.M., Lloyd, D.E., Corwin, E.M., Penaloza, M.L., Feind, R.E., Berendes, T.A., Kuo, K.-S., Welch, R.M.: The ASTER polar cloud mask. IEEE Trans. Geosci. Remote Sens. **36**(4), 1302–1312 (1998)
50. Murtagh, F., Barreto, D., Marcello, J.: Decision boundaries using Bayes factors: the case of cloud masks. IEEE Trans. Geosci. Remote Sens. **41**(12), 2952–2958 (2003)
51. Demoment, G.: Image reconstruction and restoration: overview of common estimation structures and problems. IEEE Trans. Acoust. Speech Signal Process. **37**(12), 2024–2036 (1989)
52. Banham, M.R., Katsaggelos, A.K.: Digital image restoration. IEEE Signal Process. Mag. **14**(2), 24–41 (1997)
53. Reichenbach, S.E., Koehler, D.E., Strelow, D.W.: Restoration and reconstruction of AVHRR images. IEEE Trans. Geosci. Remote Sens. **33**(4), 997–1007 (1995)
54. Petrovich Bakalov, V., Yuryevich Yerokhin, M.: Removal of uncontrollable phase distortions in synthetic aperture radar signals. IEEE Trans. Geosci. Remote Sens. **38**(3), 1298–1302 (2000)
55. Reichenbach, S.E., Li, J.: Restoration and reconstruction from overlapping images for multi-image fusion. IEEE Trans. Geosci. Remote Sens. **39**(4), 769–780 (2001)
56. Wu, Z., Liu, C.: An image reconstruction method using GPR data. IEEE Trans. Geosci. Remote Sens. **37**(1), 327–334 (1999)
57. Holben, B.N.: Characteristics of maximum-value composite images from temporal AVHRR data. Int. J. Remote Sens. **7**(11), 1417–1434 (1986)
58. Lee, S., Crawford, M.: An adaptive reconstruction system for spatially correlated multispectral multitemporal images. IEEE Trans. Geosci. Remote Sens. **29**(4), 494–508 (1991)
59. Lee, S., Crawford, M.M.: Adaptive reconstruction of sequential AVHRR imagery of Texas via dynamic compositing using an exponentially weighted polynomial function. In: IEEE International Geoscience and Remote Sensing Symposium: IGARSS'94, Surface and Atmospheric Remote Sensing: Technologies, Data Analysis and Interpretation, vol. 1, pp. 64–66 (1994)
60. Cihlar, J., Howarth, J.: Detection and removal of cloud contamination from AVHRR images. IEEE Trans. Geosci. Remote Sens. **32**(3), 583–589 (1994)
61. Choudhury, B., Tucker, C.: Satellite observed seasonal and inter-annual variation of vegetation over the Kalahari, the Great Victoria Desert, and the Great Sandy Desert: 1979–1984. Remote Sens. Environ. **23**(2), 233–241 (1987)
62. Long, D.G., Remund, Q.P., Daum, D.L.: A cloud-removal algorithm for SSM/I data. IEEE Trans. Geosci. Remote Sens. **37**(1), 54–62 (1999)
63. Gao, B.-C., Yang, P., Han, W., Li, R.-R., Wiscombe, W.J.: An algorithm using visible and 1.38-μm channels to retrieve cirrus cloud reflectances from aircraft and satellite data. IEEE Trans. Geosci. Remote Sens. **40**(8), 1659–1668 (2002)
64. Moody, E.G., King, M.D., Platnick, S., Schaaf, C.B., Gao, F.: Spatially complete global spectral surface albedos: value-added datasets derived from Terra MODIS land products. IEEE Trans. Geosci. Remote Sens. **43**(1), 144–158 (2005)
65. Tseng, D.-C., Tseng, H.-T., Chien, C.-L.: Automatic cloud removal from multi-temporal spot images. Appl. Math. Comput. **205**(2), 584–600 (2008)
66. Lin, C.-H., Tsai, P.-H., Lai, K.-H., Chen, J.-Y.: Cloud removal from multitemporal satellite images using information cloning. IEEE Trans. Geosci. Remote Sens. **51**(1), 232–241 (2013)
67. Melgani, F.: Contextual reconstruction of cloud-contaminated multitemporal multispectral images. IEEE Trans. Geosci. Remote Sens. **44**(2), 442–455 (2006)
68. Lorenzi, L., Melgani, F., Mercier, G.: Missing-area reconstruction in multispectral images under a compressive sensing perspective. IEEE Trans. Geosci. Remote Sens. **51**(7), 3998–4008 (2013)
69. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. J. R. Statis. Soc. Ser. B (Methodol.) 1–38 (1977)
70. Moon, T.K.: The expectation-maximization algorithm. IEEE Signal Process. Mag. **13**(6), 47–60 (1996)
71. Rissanen, J.: Stochastic Complexity in Statistical Enquiry. World Scientific, Singapore (1989)
72. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification. Wiley, New York (2001)

73. Kittler, J., Illingworth, J.: Minimum error thresholding. Pattern Recognit. **19**(1), 41–47 (1986)
74. Sezan, M.I.: A peak detection algorithm and its application to histogram-based image data reduction. Comput. Vis. Gr. Image Process. **49**(1), 36–51 (1990)
75. Yen, J.-C., Chang, F.-J., Chang, S.: A new criterion for automatic multilevel thresholding. IEEE Trans. Image Process. **4**(3), 370–378 (1995)
76. Shah-Hosseini, H., Safabakhsh, R.: Automatic multilevel thresholding for image segmentation by the growing time adaptive self-organizing map. IEEE Trans. Pattern Anal. Mach. Intell. **10**, 1388–1393 (2002)
77. McLachlan, G., Peel, D.: Multivariate normal mixtures. Finite Mixture Models, pp. 81–116 (2000)
78. Rissanen, J.: Complexity in Statistical Inquiry. Teaneck (1989)
79. Liang, Z., Jaszczak, R.J., Coleman, R.E.: Parameter estimation of finite mixtures using the EM algorithm and information criteria with application to medical image processing. IEEE Trans. Nucl. Sci. **39**(4), 1126–1133 (1992)
80. Donoho, D.L.: Compressed sensing. IEEE Trans. Inf. Theory **52**(4), 1289–1306 (2006)
81. Candès, E.J., Romberg, J., Tao, T.: Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. IEEE Trans. Inf. Theory **52**(2), 489–509 (2006)
82. Candes, E.J., Tao, T.: Decoding by linear programming. IEEE Trans. Inf. Theory **51**(12), 4203–4215 (2005)
83. Chen, S.S., Donoho, D.L., Saunders, M.A.: Atomic decomposition by basis pursuit. SIAM Rev. **43**(1), 129–159 (2001)
84. Bruckstein, A.M., Donoho, D.L., Elad, M.: From sparse solutions of systems of equations to sparse modeling of signals and images. SIAM Rev. **51**(1), 34–81 (2009)
85. Wang, N., Wang, Y.: An image reconstruction algorithm based on compressed sensing using conjugate gradient. In: IEEE 2010 4th International Universal Communication Symposium (IUCS), pp. 374–377 (2010)
86. Pati, Y.C., Rezaiifar, R., Krishnaprasad, P.: Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition. In: IEEE 1993 Conference Record of The Twenty-Seventh Asilomar Conference on Signals, Systems and Computers, pp. 40–44 (1993)
87. Tropp, J.A., Gilbert, A.C.: Signal recovery from random measurements via orthogonal matching pursuit. IEEE Trans. Inf. Theory **53**(12), 4655–4666 (2007)
88. Kunis, S., Rauhut, H.: Random sampling of sparse trigonometric polynomials II - orthogonal matching pursuit versus basis pursuit. Found. Comput. Math. **8**(6), 737–763 (2008)
89. Goldberg, D.E.: Genetic Algorithms in Search and Machine Learning. Addison Wesley, Reading (1989)
90. Chambers, L.D.: Practical Handbook of Genetic Algorithms: Complex Coding Systems, vol. 3. CRC Press, Boca Raton (1998)
91. Deb, K.: Multi-objective optimization using evolutionary algorithms, vol. 16. Wiley, New York (2001)
92. Zitzler, E., Laumanns, M., Thiele, L., Zitzler, E., Zitzler, E., Thiele, L., Thiele, L.: SPEA2: Improving the Strength Pareto Evolutionary Algorithm (2001)
93. Knowles, J., Corne, D.: The pareto archived evolution strategy: a new baseline algorithm for pareto multiobjective optimisation. In: IEEE Proceedings of the 1999 Congress on Evolutionary Computation, CEC 99, vol. 1 (1999)
94. Srinivas, N., Deb, K.: Muiltiobjective optimization using nondominated sorting in genetic algorithms. Evol. Comput. **2**(3), 221–248 (1994)
95. Ghoggali, N., Melgani, F., Bazi, Y.: A multiobjective genetic SVM approach for classification problems with limited training samples. IEEE Trans. Geosci. Remote Sens. **47**(6), 1707–1718 (2009)
96. Ghoggali, N., Melgani, F.: Genetic SVM approach to semisupervised multitemporal classification. IEEE Geosci. Remote Sens. Lett. **5**(2), 212–216 (2008)

97. Pasolli, E., Melgani, F., Donelli, M.: Automatic analysis of GPR images: a pattern-recognition approach. IEEE Trans. Geosci. Remote Sens. **47**(7), 2206–2217 (2009)
98. Liu, C.-C.: Processing of FORMOSAT-2 daily revisit imagery for site surveillance. IEEE Trans. Geosci. Remote Sens. **44**(11), 3206–3214 (2006)
99. Baudoin, A.: Mission analysis for spot 5. In: IEEE International Geoscience and Remote Sensing Symposium, IGARSS'93, Better Understanding of Earth Environment., p. 1084 (1993)
100. Jain, A.K.: Fundamentals of Digital Image Processing. Prentice-Hall Inc., Englewood Cliffs (1989)

Chapter 10
Advances in Kernel Machines for Image Classification and Biophysical Parameter Retrieval

Devis Tuia, Michele Volpi, Jochem Verrelst and Gustau Camps-Valls

Abstract Remote sensing data analysis is knowing an unprecedented upswing fostered by the activities of the public and private sectors of geospatial and environmental data analysis. Modern imaging sensors offer the necessary spatial and spectral information to tackle a wide range problems through Earth Observation, such as land cover and use updating, urban dynamics, or vegetation and crop monitoring. In the upcoming years even richer information will be available: more sophisticated hyperspectral sensors with high spectral resolution, multispectral sensors with sub-metric spatial detail or drones that can be deployed in very short time lapses. Besides such opportunities, these new and wealthy information sources also come with a price: the analysts are confronted with data showing large and complex feature characteristics. To deal with these new challenges, kernel methods have emerged as a valid, robust and successful framework. The intrinsic regularization implemented in these methods and their low sensitivity to data dimensionality make them natural candidates to solve current remote sensing problems. The flexibility offered by kernel methods allows us to treat heavily nonlinear tasks with elegant methodologies, while still using linear algebra. In the last decade, kernel methods in general, and support vector machines for classification and Gaussian processes for regression in particular, have become standard tools for geospatial data analysis. In this chapter, we first review the main concepts about kernel methods and their use in remote sensing. Then, we review

D. Tuia (✉)
Laboratory of GeoInformation Science and Remote Sensing,
Wageningen University and Research, Droevendaalsesteeg 3, 6708 PB
Wageningen, The Netherlands
e-mail: devis.tuia@wur.nl

M. Volpi
Department of Geography, University of Zurich, Winterthurerstrasse 190,
8057 Zurich, Switzerland
e-mail: michele.volpi@geo.uzh.ch

J. Verrelst · G. Camps-Valls
Image Processing Laboratory, Universitat de València, Valencia, Spain
e-mail: jochem.verrelst@uv.es

G. Camps-Valls
e-mail: gustau.camps@uv.es

G. Moser and J. Zerubia (eds.), *Mathematical Models for Remote Sensing Image Processing*, Signals and Communication Technology,
https://doi.org/10.1007/978-3-319-66330-2_10

examples of kernel methods for remote sensing image classification and biophysical parameter retrieval.

10.1 Introduction

Analyzing remote sensing data has become more challenging in recent years. As mentioned in previous chapters, the volume of data to be processed has increased considerably. Secondly, the resolution (both in spectral and spatial terms) has improved, thus allowing to study more interesting problems, but also of higher complexity. Third, the diversity of the data has also increased, in the sense that several sensors are nowadays available to the general public, each one with its own characteristics and (dis)advantages. In the area of modern remote sensing, there is a need for new processing methods capable of handling data from different sensors, with high complexity and coming at a high spectral, spatial and temporal resolutions.

Among the different families of statistical methods that can be considered to process remote sensing data, kernel methods [1] have gained popularity in the last decades. Kernel methods are theoretically sound, reduce to linear operators while providing nonlinear solutions to the problem at hand, and cast processing problems in terms of the estimation of similarities between data samples. In this sense, kernel methods are generally non-parametric methods, i.e. they learn the dependencies between the inputs and the outputs in a data-driven way, without assuming a parametric model generating the data.

The fact that kernel methods (and most generally non-parametric methods) rely only on the labeled data might lead to problems, since, on one hand, learning from data would imply the necessity of large data sets, while on the other hand the reduced availability of labeled samples implies the risk of missing the general structure of data (since we stay too close to these training data). To leverage these critical issues, kernel methods employ regularization strategies limiting the complexity of the resulting models: by favouring simpler models over more complex ones relying massively on the training data, kernel methods achieve remarkable generalization ability [1]. This is even further increased when enforcing other types of regularity or sources of prior knowledge about the data at hand: priors such as sparsity [4] or smoothness [5] have been employed successfully in remote sensing image processing and will also be reviewed in this chapter.

The first task we will review in this chapter is pixel classification, i.e. the act of classifying every pixel (or region) of the image in a specific semantic class: in this task, kernel methods have shown excellent generalization abilities [6–9], robustness to small sample scenarios [10, 11] and flexibility related to the possibility of designing kernel functions specific to the problem [12, 13]. Support vector machines are

nowadays one of the most used methods in remote sensing and have imposed themselves as the golden standard for land cover and land use classification [14–17] (see also the various applications of SVM classifiers in previous chapters). In Section 10.3 we will review some recent advances in multi-sensor classification of remote sensing data, and describe in detail two approaches: a multi-kernel approach to integrate multi-temporal sequences, and a manifold alignment approach to perform domain adaptation, i.e. adapt a model based on existing training samples to predict effectively the classes in a new image acquisition (thus acquired under different illumination, seasonal and atmospheric conditions).

The second task we will review in this chapter is the estimation of biophysical parameters by means of statistical machine learning. Spatio-temporally explicit and quantitative retrieval of the characteristics of the surface of the Earth or of its atmosphere have become a requirement in a variety of Earth observation applications. Optical sensors mounted on-board Earth observation (EO) satellites are being endowed with high temporal, spectral and spatial resolutions, and thus enable the retrieval and monitoring of climate and bio-geophysical variables [18, 19]. With the super-spectral Copernicus Sentinel-2 (S2) [20] and the forthcoming Sentinel-3 missions [21], among other planned space missions, an unprecedented data stream for land, ocean and atmosphere monitoring will soon become available to a diverse user community. Such vast data streams require enhanced processing techniques and statistical inference methods might play an important role in this area of research. Over the last few decades a wide diversity of bio-geophysical retrieval methods have been developed, but only a few of them made it into operational processing chains. Essentially, we may find two main approaches to the inverse problem of estimating biophysical parameters from spectra: *parametric physically-based models* and *non-parametric statistical models*. Lately, machine learning has attained outstanding results in the estimation of climate variables and related bio-geophysical parameters at local and global scales [22]. For example, leaf area index (LAI) [23] and Gross Primary Production (GPP) [24, 25] are currently derived with neural networks, while multiple regression is used for retrieving biomass [26] or sun-induced fluorescence [27]. Support vector methods were also proposed to study ocean chlorophyll [28] and vegetation parameters [29–31]. The family of Bayesian non-parametrics, and of Gaussian processes in particular [3], have been payed wide attention in the last years in remote sensing data analysis, especially for tasks of vegetation properties estimation [32]. We will review the main developments in GPs for EO data analysis in this chapter (Section 10.4). We review new algorithms that respect the signal characteristics, that provide feature rankings automatically, and that allow applicability of associated uncertainty intervals to transport GP models in space and time.

10.2 Introduction to Kernel Methods

In this section, we review the main properties of kernel methods. By drawing from linear algebra and functional analysis we discuss valid operations on kernels functions

[33, 34]. Furthermore, we present two methods in details: support vector machines for classification and Gaussian processes for regression.

10.2.1 Feature Maps and Kernels

Kernel methods are learning algorithms that extrapolate rules for inference by learning the structure between training examples. The rules are based on similarities between samples. They rely on and extend the well known theory of linear algorithms, so that nonlinear models can be achieved by using linear algebra.

The simplest solution to make linear algorithms deal with nonlinear problems is to map the *input* samples $S = \{\mathbf{x}_i \in \mathbb{R}^d\}_{i=1}^n \in \mathscr{X}$ into a higher dimensional Hilbert space $\mathscr{H}$, the *feature space*, endowed with the dot product operation. The solution obtained by learning the linear model in this space will be nonlinear with respect to the original input space $\mathscr{X}$. The nature of $\mathscr{X}$ is unimportant, as long as we can define a mapping function $\boldsymbol{\phi} : \mathscr{X} \to \mathscr{H}, \mathbf{x} \mapsto \boldsymbol{\phi}(\mathbf{x})$. This mapping step is very important, since it allows to build valid kernel functions also over non-conventional input spaces such as strings or histograms: the similarity among data instances will be measured by the inner product $\langle \cdot, \cdot \rangle_{\mathscr{H}}$ in $\mathscr{H}$.

However, computing explicitly the optimal $\boldsymbol{\phi}(\cdot)$ guaranteeing low generalization error is a computationally very expensive task, since the whole transformation must be estimated from the data. Kernel methods circumvent this problem elegantly by defining a *kernel* function returning directly the value of the pairwise inner product in $\mathscr{H}$ by only taking as argument the samples in their original input space $K : \mathscr{X} \times \mathscr{X} \to \mathbb{R}$, such that $(\mathbf{x}, \mathbf{x}') \mapsto K(\mathbf{x}, \mathbf{x}')$. A *kernel function* is defined to be:

$$K(\mathbf{x}, \mathbf{x}') = \langle \boldsymbol{\phi}(\mathbf{x}), \boldsymbol{\phi}(\mathbf{x}') \rangle_{\mathscr{H}}. \tag{10.1}$$

10.2.2 Positive Definite Kernels and the Kernel Trick

Using the kernel function in Eq. (10.1) in all sample pairwise combinations, we can construct a square symmetric *kernel matrix* of real numbers, $\mathbf{K} \in \mathbb{R}^{n \times n}$, which contains the similarity between all available data points in its entries $K(\mathbf{x}_i, \mathbf{x}_j)$. This matrix is *positive semi-definite* if, for any two scalars $c_i, c_j \in \mathbb{R}$, $\sum_{ij} c_i c_j K(\mathbf{x}_i, \mathbf{x}_j) \geq 0$. If the equality occurs only at $c_i = c_j = \ldots = 0$, the kernel matrix is strictly positive definite. A kernel function is valid if and only if the associated kernel matrix $\mathbf{K}$ is positive semi-definite.

Theorem 1 *There exists a Hilbert space* $\mathscr{H}$ *and a feature map* $\boldsymbol{\phi} : \mathscr{X} \rightarrow \mathscr{H}$ *for all* $(\mathbf{x}, \mathbf{x}') \in \mathscr{X}$ *so that* $K(\mathbf{x}, \mathbf{x}') = \langle \boldsymbol{\phi}(\mathbf{x}), \boldsymbol{\phi}(\mathbf{x}') \rangle_{\mathscr{H}}$ if and only if $K(\mathbf{x}, \mathbf{x}')$ *gives rise to a positive definite* kernel matrix $\mathbf{K} \in \mathbb{R}^{n \times n}$.

The above definition gives rise to the *kernel trick* [35]: Any algorithm that depends on data samples only in the form of dot products can be "kernelized" by replacing the inner products with valid kernel functions.

Note that many algorithms not explicitly depending on dot products between data can be reformulated so that kernels can be applied, e.g., principal component analysis and least squares regression, to name a few. By doing so, the algorithm will run by implicitly depending on inner products in $\mathscr{H}$ without needing the explicit form of $\boldsymbol{\phi}(\cdot)$.

10.2.3 Operations with Kernels

The feature space in which data samples are mapped is endowed with a dot product and basic algebraic operations. Thus, there exist a set of basic operations that can be applied in $\mathscr{H}$ by means of kernels.

Translation. A translation by a vector $\boldsymbol{\Gamma} \in \mathscr{H}$ corresponds to the modified mapping function $\tilde{\boldsymbol{\phi}}(\mathbf{x}) = \boldsymbol{\phi}(\mathbf{x}) + \boldsymbol{\Gamma}$. If $\boldsymbol{\Gamma}$ lies in the span of the mapped data samples, the dot product between translated mapped samples can be readily computed as:

$$\langle \tilde{\boldsymbol{\phi}}(\mathbf{x}), \tilde{\boldsymbol{\phi}}(\mathbf{x}') \rangle_{\mathscr{H}} = \langle \boldsymbol{\phi}(\mathbf{x}) + \boldsymbol{\Gamma}, \boldsymbol{\phi}(\mathbf{x}) + \boldsymbol{\Gamma} \rangle_{\mathscr{H}}$$

Centering. The fact that translations exist in $\mathscr{H}$ allows us to center data directly in the feature space. The mean of a data set in $\mathscr{H}$ can be computed as $\boldsymbol{\phi}_{\mu} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{\phi}(\mathbf{x}_i)$, which lies in the span of the $\boldsymbol{\phi}(\mathbf{x})$, since the mean is a convex combination of data samples. Centered mappings $\bar{\boldsymbol{\phi}}(\mathbf{x})$ correspond to a translation with the data mean: $\bar{\boldsymbol{\phi}}(\mathbf{x}) = \boldsymbol{\phi}(\mathbf{x}) - \boldsymbol{\phi}_{\mu}$. Thus, the dot product between centered data in $\mathscr{H}$ is:

$$\begin{aligned} \langle \bar{\boldsymbol{\phi}}(\mathbf{x}_i), \bar{\boldsymbol{\phi}}(\mathbf{x}_j) \rangle_{\mathscr{H}} &= \langle \boldsymbol{\phi}(\mathbf{x}_i) - \boldsymbol{\phi}_{\mu}, \boldsymbol{\phi}(\mathbf{x}_j) - \boldsymbol{\phi}_{\mu} \rangle_{\mathscr{H}} \\ &= K(\mathbf{x}_i, \mathbf{x}_j) - \frac{2}{n} \sum_{i=1}^{n} K(\mathbf{x}_i, \mathbf{x}_j) + \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} K(\mathbf{x}_i, \mathbf{x}_j) \end{aligned} \tag{10.2}$$

Computing Distances. Another direct application of the translation operation allows to compute distances in the *feature space*. It is evaluated entirely in terms of kernel functions:

$$d(\mathbf{x}, \mathbf{x}')_{\mathscr{H}} = \|\boldsymbol{\phi}(\mathbf{x}) - \boldsymbol{\phi}(\mathbf{x}')\|_{\mathscr{H}} = \sqrt{K(\mathbf{x}, \mathbf{x}) + K(\mathbf{x}', \mathbf{x}') - 2K(\mathbf{x}, \mathbf{x}')}$$

Subspace Projections. Geometrical operations in $\mathscr{H}$ are fully defined. Consequently, given two vectors $\boldsymbol{\Psi}, \boldsymbol{\Gamma} \in \mathscr{H}$, the projection of $\boldsymbol{\Psi}$ onto the subspace spanned by $\boldsymbol{\Gamma}$ is

$$\boldsymbol{\Psi}' = \frac{\langle \boldsymbol{\Gamma}, \boldsymbol{\Psi} \rangle_{\mathscr{H}}}{\langle \boldsymbol{\Gamma}, \boldsymbol{\Gamma} \rangle_{\mathscr{H}}} \boldsymbol{\Gamma}.$$

By applying properties seen above, one can express the dot product between projections Ψ' entirely by kernel evaluations.

Normalization. From the definition of subspace projections, it results that computing dot products between normalized mapped samples $\hat{\boldsymbol{\phi}}(\mathbf{x})$ to unit norm becomes:

$$\langle \hat{\boldsymbol{\phi}}(\mathbf{x}), \hat{\boldsymbol{\phi}}(\mathbf{x}') \rangle_{\mathscr{H}} = \left\langle \frac{\boldsymbol{\phi}(\mathbf{x})}{\|\boldsymbol{\phi}(\mathbf{x})\|}, \frac{\boldsymbol{\phi}(\mathbf{x}')}{\|\boldsymbol{\phi}(\mathbf{x}')\|} \right\rangle = \frac{K(\mathbf{x}, \mathbf{x}')}{\sqrt{K(\mathbf{x}, \mathbf{x}) K(\mathbf{x}', \mathbf{x}')}}. \tag{10.3}$$

10.2.4 Kernels Functions

In kernel methods, the choice of the kernel function and the corresponding hyperparameters (model selection) plays a very important role. The kernel implements a functional prior, in the sense that the kernel function and its hyperparameter set must fit the data in order to encode their relationships. In other words, the choice of the kernel function implicitly specifies the *form* of the mapping function $\boldsymbol{\phi}(\cdot)$ *and* the dot product (i.e., the similarity) between mappings.

In general, there exist three widely used kernel functions: linear, polynomial, and radial basis functions (RBF). The three most famous instances are as follows:

$$\text{Linear kernel: } K(\mathbf{x}, \mathbf{x}') = \langle \mathbf{x}, \mathbf{x}' \rangle \tag{10.4}$$

$$\text{Polynomial kernel: } K(\mathbf{x}, \mathbf{x}') = \left(\langle \mathbf{x}, \mathbf{x}' \rangle + 1 \right)^p, \; p \in \mathbb{N}^+ \tag{10.5}$$

$$\text{Gaussian RBF kernel: } K(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{1}{2\sigma^2} \|\mathbf{x} - \mathbf{x}'\|^2 \right), \; \sigma \in \mathbb{R}^+, \tag{10.6}$$

The hyperparameters b, p, σ above specify a specific kernel among the corresponding function family. Running the kernelized algorithm with the linear kernel corresponds, up to some numerical differences, to running the original algorithm in the original space. For the polynomial, p correspond to the degree of the polynomial expansion (thus a mapping of $p+1$ dimensions) and with the 1 allows to account for different monomials of the power. Note that linear kernels are a particular instance of polynomial kernels with $p = 1$ and no bias term.

For the Gaussian RBF, σ controls the bandwidth. This hyperparameter controls the degree of locality of the Gaussian RBF kernel, weighting the distance between sample pairs. For this reason, it belongs to the family of local kernels. The ratio

inside the exponential grows towards large negative numbers for very small σ (and thus the exponential tends towards 0). On the opposite, for very large σ, the ratio tends towards 0 and the exponential towards 1. When σ becomes arbitrarily large, the RBF kernel roughly behave like a linear kernel under some mild assumptions. This is readily seen by expanding the Euclidean norm in the exponential. By Taylor series expansion, the Gaussian RBF kernel approximates to a polynomial of infinite degree, corresponding to $\dim(\mathscr{H}) \to \infty$. This kernel function is probably the most widely used, as it offers an easy interpretation (local similarity with bounded output) and it tends to outperform other basic kernel functions in practice.

The problem of finding an explicit feature map is now reduced to select a kernel function and to tune the corresponding hyperparameters.

These tasks are purely data dependent and can be performed by model selection strategies (cross-validation) or by optimization routines (such as the multiple kernel learning, see Sect. 10.3 for SVM and Sect. 10.2.8 for GPR).

10.2.5 Kernel Combinations

One of the most interesting aspects of kernel functions is that they can also be *derived* from other kernel functions. As illustrated by linear algebra and function analysis key properties [33, 34], new kernels can be built by taking advantage of a set of rules, which define the validity of some operations *between* kernels [35]. Let us now assume $K_1(\cdot,\cdot)$ and $K_2(\cdot,\cdot)$ two positive definite kernel functions on $\mathscr{X} \times \mathscr{X}$, $\mathbf{A}$ a symmetric positive semidefinite matrix (e.g., a covariance matrix or an inducing metric), $M(\cdot,\cdot)$ a metric function fulfilling the triangle inequality and finally $f(\cdot)$ any continuous function. The rules are as follows:

$$K(\mathbf{x},\mathbf{x}') = K_1(\mathbf{x},\mathbf{x}') + K_2(\mathbf{x},\mathbf{x}') \tag{10.7}$$
$$K(\mathbf{x},\mathbf{x}') = \mu K(\mathbf{x},\mathbf{x}'),\ \mu > 0 \tag{10.8}$$
$$K(\mathbf{x},\mathbf{x}') = K_1(\mathbf{x},\mathbf{x}') \cdot K_2(\mathbf{x},\mathbf{x}') \tag{10.9}$$
$$K(\mathbf{x},\mathbf{x}') = \mathbf{x}^\top \mathbf{A}\mathbf{x}' \tag{10.10}$$
$$K(\mathbf{x},\mathbf{x}') = \exp(-M(\mathbf{x},\mathbf{x}')) \tag{10.11}$$
$$K(\mathbf{x},\mathbf{x}') = K(f(\mathbf{x}), f(\mathbf{x}')) \tag{10.12}$$

It results that, by choosing appropriate kernel functions and combining them we can develop new similarity measures that better encode data/problem characteristics, making our priors closer to data modeling needs.

For instance, let assume that we are dealing with a multi-source classification problem (cf. Chap. 7). We know from the problem that some sources of data are generally more important than others, but even less informative sources give important information for some classes. We also know that one important source of data is very noisy and we would like to utilize a smoothed version of the image without knowing which smoothing function is best. By the above rules, we can build a kernel that accounts for these aspects simultaneously: we can weight (by a convex combination) the kernel function for each data source together with the different smoothing for the noisy data source. By learning these weights from the data (e.g., by minimum error cross-validation) we end up in having a kernel matrix optimized for the problem at hand. A detailed example of this reasoning will be given in Sect. 10.3, and a Bayesian perspective will be presented in Sect. 10.4.

10.2.6 A Note on the Kernel Metric

Kernel methods may appear elusive because the mapping $\boldsymbol{\phi}$ is not explicitly defined, and the vector coordinates in the new feature spaces are not accessible. However, the framework allows to compute distances, angles, displacements, averages, and covariances implicitly in $\mathscr{H}$ from the available data [36].

In addition, and very importantly, we show here that one can compute the metric associated to the used kernel.

For any positive definite kernel, we assume that the mapped data in $\mathscr{H}$ are distributed in a surface $\mathscr{S}$ smooth enough to be considered a Riemannian manifold [37]. The line element of $\mathscr{S}$ can be expressed as:

$$ds^2 = g_{ab}\mathrm{d}\boldsymbol{\phi}^a(\mathbf{x})\mathrm{d}\boldsymbol{\phi}^b(\mathbf{x}) = g_{\mu\nu}\mathrm{d}x^\mu \mathrm{d}x^\nu,$$

where superscripts a and b correspond to the vector space $\mathscr{H}$, $g_{\mu\nu}$ is the induced metric, and the surface $\mathscr{S}$ is parametrized by x^μ. Note that Einstein's summation convention over repeated indices is used. Computing the components of the (symmetric) metric tensor only needs the kernel function:

$$g_{\mu\nu} = (1/2)\partial_{x_\mu}\partial_{x_\nu}K(\mathbf{x},\mathbf{x}) - \{\partial_{x'_\mu}\partial_{x'_\nu}K(\mathbf{x},\mathbf{x}')\}_{\mathbf{x}'=\mathbf{x}}. \tag{10.13}$$

For the RBF kernel with a given σ parameter, this metric tensor becomes flat, $g_{\mu\nu} = \delta_{\mu\nu}/\sigma^2$, and the squared geodesic distance between $\boldsymbol{\phi}(\mathbf{x})$ and $\boldsymbol{\phi}(\mathbf{x}')$ becomes:

$$\|\boldsymbol{\phi}(\mathbf{x}) - \boldsymbol{\phi}(\mathbf{x}')\|^2_{\mathscr{H}} = 2\left(1 - \exp\left(-\frac{\|\mathbf{x}-\mathbf{x}'\|^2_{\mathscr{X}}}{2\sigma^2}\right)\right) = 2(1 - K(\mathbf{x},\mathbf{x}')). \tag{10.14}$$

Note that the metric solely depends on the original data points (e.g., spectra) yet computed implicitly in a higher dimensional feature space $\mathscr{H}$, whose notion of distance is controlled by the parameter σ: the larger σ the smoother (linear) is the space. Actually, $\sigma \to \infty$ reduces the RBF kernel to approximately compute the Euclidean distance between vectors, which reduces the metric tensor to $g_{\mu v} = 1$.

10.2.7 Support Vector Machine for Classification

Support Vector Machine is a linear classifier maximizing a margin between data instances belonging to different classes. This stems directly from an implementation of the structural risk minimization, as dictated by the statistical learning theory [38]. In few words, SVM minimizes the regularized empirical loss on training data, by finding a solution jointly minimizing the error on training data (empirical risk) *and* selecting the model with the optimal complexity (regularization loss).

The SVM is defined as follows. We dispose of a set of data-label pairs $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$, where $\mathbf{x}_i \in \mathbb{R}^d$ and $y_i \in \{-1, +1\}$. To obtain a nonlinear SVM, we solve the problem in $\mathscr{H}$ by applying to the data samples the mapping function $\boldsymbol{\phi}(\cdot)$. Then, the SVM primal optimization problem is:

$$\min_{\mathbf{w},\xi_i,b} \left\{ \frac{1}{2}\|\mathbf{w}\|^2 + C \sum_{i=1}^{n} \xi_i \right\} \tag{10.15}$$

constrained to:

$$y_i(\langle \boldsymbol{\phi}(\mathbf{x}_i), \mathbf{w} \rangle_{\mathscr{H}} + b) \geq 1 - \xi_i \qquad \forall i = 1, \ldots, n \tag{10.16}$$

$$\xi_i \geq 0 \qquad \forall i = 1, \ldots, n \tag{10.17}$$

The optimal $\mathbf{w}$ and b define the (primal) model, that is simply a linear classifier in the feature space for some test sample $\mathbf{x}_*$ as $y_* = \text{sign}\,(\langle \mathbf{x}_*, \mathbf{w} \rangle + b)$. The ξ_i are positive slack variables allowing the optimization to deal with training errors, typically caused by noise and data errors which do not allow to learn a perfect separation. The regularization hyperparameter C controls the capacity of the SVM, by penalizing (small C) or encouraging (large C) complex models, defined by the smoothness of the weights $\mathbf{w}$.

As discussed above, the form of the optimal mapping $\boldsymbol{\phi}(\cdot)$ is impossible to be explicitly computed. We know that we can circumvent the problem by applying the kernel trick and select the appropriate mappings by only tuning kernel hyperparameters. To obtain a formulation only dependent on dot products between mapped samples the above primal problem is often reformulated via its dual counterpart.

It is obtained by equating partial derivatives of the model to 0, and replacing the optimal points in Eq. (10.15). The dual optimization corresponds to the following maximization with respect to the dual Lagrange multipliers $\boldsymbol{\alpha}$ [35]:

$$\max_{\boldsymbol{\alpha}} \left\{ \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) \right\} \tag{10.18}$$

subject to:

$$0 \leq \alpha_i \leq C \qquad \forall i = 1, \ldots, n \tag{10.19}$$

$$\sum_{i=1}^{n} y_i \alpha_i = 0 \qquad \forall i = 1, \ldots, n \tag{10.20}$$

Once the optimal model parameter α_i are obtained, the final (dual) decision on some test example $\mathbf{x}_*$ can be made by taking the sign of the linear regression between labels defined by $\mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \boldsymbol{\phi}(\mathbf{x}_i)$ and $b = 1/k \sum_{i=1}^{k} (y_i - \langle \boldsymbol{\phi}(\mathbf{x}_i), \mathbf{w} \rangle)$. For the latter, k is the number of *unbounded* Lagrange multipliers ($0 < \alpha_i < C$) and the sum is taken over the corresponding samples. The dual solution for a test sample $\mathbf{x}_*$ is then formulated as:

$$y_* = \operatorname{sgn}\left(\sum_{i=1}^{n} y_i \alpha_i K(\mathbf{x}_i, \mathbf{x}_*) + b \right) \tag{10.21}$$

Note that inferring a new label with a SVM requires that the training samples with an associated $\alpha_i \neq 0$ are stored in memory, in order to compute the kernel matrix at test time. There exist some ways to reduce the computational load of kernel methods, such as kernel decompositions based on Nyström approximations [39] or on Landmark selection [40]. Other strategies suggest to solve directly the primal problem using for instance approximate mappings, such as random feature sinks [41].

Standard SVM is binary in nature. Multi-class extensions to solve c-class problems usually train several SVM, one per binary problem, by two main splitting strategies: one against one and one against all. The former trains $c(c-1)/2$ SVM while the latter only c. SVM directly solving a multi-class optimization problem are usually more efficient when solving the primal problem (scaling linearly with the number of examples), but become more expensive when solving the kernel counterpart. A typical instance of this problem can be easily formulated by structured SVM, known as the Crammer and Singer SVM [42].

10.2.8 Gaussian Processes for Regression

Gaussian Processes (GPs) belong to the family of Bayesian nonparametric methods [3]. Under a pure discriminative perspective, they can be actually seen as a kernel method performing least squares regression in a reproducing kernel Hilbert space (RKHS): dual weights can be estimated easily involving the inverse of the (regularized) training kernel matrix, and predictions for new test data involve the computation of a kernel matrix containing the similarities between training and test data. In the GPs literature, the *kernel matrix* is called the *covariance matrix*.

GPs find a regression model of the type $f(\mathbf{x}) = \langle \mathbf{w}, \boldsymbol{\phi}(\mathbf{x}) \rangle_{\mathscr{H}} + \varepsilon$, where ε is a zero mean σ_r^2 standard deviation Gaussian white noise, by defining a distribution over the functions, *i.e.* a Gaussian Process, $f(\mathbf{x}) \sim \text{GP}(m(\mathbf{x}), K(\mathbf{x}, \mathbf{x}'))$ fully specified by a mean function $m(\mathbf{x}) = E[f(\mathbf{x})]$ and covariance (kernel) $K(\mathbf{x}, \mathbf{x}') = E[(f(\mathbf{x}) - E[f(\mathbf{x})])\,(f(\mathbf{x}) - E[f(\mathbf{x})])]$. The mean function, for the sake of simplicity will be assumed to be zero. This would not limit the modeling power, as long as the data is centered.

We dispose of a labeled training data set with outputs (real numbers) $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$, where $\mathbf{x}_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$. To infer the GP model, we only have to compute the kernel matrix $\mathbf{K}$ which now fully specifies the distribution over functions. Practically, this distribution models the *outputs* of each sample from the GP. The covariance matrix $\mathbf{K}$ specifies that estimated training set outputs $\mathbf{f}(\mathbf{x}) \sim \mathscr{N}(\mathbf{0}, \mathbf{K})$. Thus, the form of this covariance function and in particular of its hyperparameters will define the form of variation across functions.

Let us define the vectorized training labels as $\mathbf{y} = [y_1, \ldots, y_n]^\top$, the evaluation of kernel between training and test data $\mathbf{x}_*$ as $\mathbf{k}_* = [K(\mathbf{x}_1, \mathbf{x}_*), \ldots, K(\mathbf{x}_n, \mathbf{x}_*)]^\top$, and the one between test samples as $k_{**} = K(\mathbf{x}_*, \mathbf{x}_*)$. All the kernel values can be seen as specific entries from a joint train-test kernel matrix. The output values are distributed according to:

$$\begin{pmatrix} \mathbf{y} \\ f(\mathbf{x}_*) \end{pmatrix} \sim \mathscr{N}\left(\mathbf{0}, \begin{pmatrix} \mathbf{K} + \sigma_r^2 \mathbf{I} & \mathbf{k}_* \\ \mathbf{k}_*^\top & k_{**} \end{pmatrix}\right) \tag{10.22}$$

To predict the output $f(\mathbf{x}_*)$ we derive the posterior distribution of the outputs by conditioning on training data, given the test sample, as

$$f(\mathbf{x}_*)|\mathbf{y}, \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}, \mathbf{x}_* \sim \mathscr{N}\left(\mathbf{k}_*(\mathbf{K} + \sigma^2\mathbf{I})^{-1}\mathbf{y}, k_{**} - \mathbf{k}_*(\mathbf{K} + \sigma^2\mathbf{I})^{-1}\mathbf{k}_*\right).$$

Therefore, predictions from a GP are simply the mean of this distribution conditioned on the observations, with an associated uncertainty given by the standard deviation of this conditional distribution.

To train the GP, we, therefore, have to optimize two sets of hyperparameters: the kernel function and the noise variance. Both can involve the optimization of one or more hyperparameters, depending on the kernel function and on the noise model. We will assume that noise is *homoscedastic* (i.e. roughly speaking the noise variance is independent of the input sample).

An alternative optimization procedure to cross-validation to choose the hyperparameters consists of using gradient descent over the negative log marginal likelihood. Actually, it is possible to obtain the likelihood of the output data by marginalizing the GP over function values, hence obtaining:

$$-\log p(\mathbf{y}|\mathbf{x},\boldsymbol{\theta}) = \frac{1}{2}\mathbf{y}^\top(\mathbf{K}+\sigma_n^2\mathbf{I})^{-1}\mathbf{y} + \frac{1}{2}\log\left(\det(\mathbf{K}+\sigma_n^2\,\mathbf{I})\right) - \frac{n}{2}\log(2\pi).$$

Then, the gradient updates are computed by taking partial derivatives of the hyperparameters $\boldsymbol{\theta}$:

$$\begin{aligned}\frac{\partial \log p(\mathbf{y}|\mathbf{x},\boldsymbol{\theta})}{\partial\theta_j} &= \frac{1}{2}\mathbf{y}^\top\mathbf{K}_r^{-1}\frac{\partial\mathbf{K}_r}{\partial\theta_j}\mathbf{K}_r^{-1}\mathbf{y} - \frac{1}{2}\mathrm{Tr}\left\{\mathbf{K}_r^{-1}\frac{\partial\mathbf{K}_r}{\partial\theta_j}\right\} \\ &= \frac{1}{2}\mathrm{Tr}\left\{(\boldsymbol{\alpha}\boldsymbol{\alpha}^\top - \mathbf{K}_r)^{-1}\frac{\partial\mathbf{K}_r}{\partial\theta_j}\right\},\end{aligned} \tag{10.23}$$

where $\boldsymbol{\alpha} = \mathbf{K}_r^{-1}\mathbf{y}$ and $\mathbf{K}_r = \mathbf{K} + \sigma_r^2\mathbf{I}$, which are computed just once. The above strategy can be used in general conjugate gradient-based optimization routines.

10.3 Multi-modal Data Classification

In this section, we focus on the challenging problem of multimodal remote sensing data classification [43]. Therefore, we will consider cases, where we perform image classification (or semantic classification/labeling) to provide a thematic class to each sample considered. The samples might be, for instance, pixels, regions, voxels or 3D points in a point cloud.

We will study two frameworks that have been recently applied in remote sensing image processing: multiple kernels classification and manifold alignment. We will provide examples on high and very high resolution image classification. We will then conclude by briefly reviewing recent emerging trends.

10.3.1 Multi-source Pixel Classification with Multiple Kernels

As they have been defined in Sect. 10.2, kernels are functions representing similarity between samples. If we focus on the definition given in Eq. (10.1), a kernel is evaluated with the same metric over the whole input space. Taking the example of the Gaussian kernel, $K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-||\mathbf{x}_i - \mathbf{x}_j||^2/(2\sigma^2))$, we can note that the same single bandwidth parameter, σ, is applied to all the features. This means that we assume that all the data distribution is isotropic, and that each feature deserves the same importance in the computation of the final similarity between samples. The metric used is therefore flat as seen in Sect. 10.2.6.

This can be sub-optimal, for example when some of the features available are noisy or strongly correlated between each other: in the first case, we would like their contribution to be minimal (i.e. a σ as small as possible), while in the latter, we would like to have a larger σ parameter, but also to be able to exclude the redundant features and account for the useful information only once. Note that this is independent from the data normalization: even if features are normalized to share the same range, one feature could be more discriminative than another one, requiring more weight in the computation of the similarity. To achieve these objectives simultaneously, we can take advantage of the kernel summation property stated in Eq. (10.7).

> We can build a single kernel per feature (or group of features) and then combine all the kernels into a new one by linear combination. Learning such a combination of kernel functions is known in the literature as *multiple kernels learning* (MKL [44]). Such a solution is more flexible and would directly account for data relationships in the model, and will also perform implicit feature selection.

MKL has been recently successfully applied in remote sensing data processing tasks, going from multisource image classification [45], to unmixing [46, 47], SAR segmentation [48] feature selection for hyperspectral image classification [49], or to combine spatio-spectral indices [50]. Sparse selection [51] and discriminative feature selection [52] were also considered. In [53], authors obtain the kernel by repeated nonlinear mappings. MKL was also combined with other machine learning frameworks such as active learning [54] or domain adaptation [55], thus going beyond the simple framework of kernel construction for classification. In the next section, we review the main ingredients of MKL.

10.3.1.1 From Composite to Multiple Kernels

The idea of combining kernels referring to different sources of data was previously explored in the *composite kernels* framework [12]: Combining the kernel summation (10.7) and kernel scaling (10.8) properties (see Sect. 10.2.5), one can formulate a valid

kernel as:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \mu_1 K_1(\mathbf{x}_i, \mathbf{x}_j) + \mu_2 K_2(\mathbf{x}_i, \mathbf{x}_j),$$

and tune the importance of the kernels involved by tuning the scalings $\mu_{1,2} > 0$. This idea was exploited for spatio-spectral classification originally in [12] for hyperspectral image classification, and further exploited for VHR [13, 56], multitemporal classification [14], and semisupervised kernel deformation [57]. Composite kernels work well in practice, as they provide an intuitive way of trade-off the importance of the different feature sets used to compute each kernel, or even to discover the best kernel function with a fixed feature set. However, they are not suitable for cases involving more than a few kernels, since the tuning of a set of μ weights heuristically (e.g., cross-validation) would become computationally expensive. The MKL framework answers to this call, as it aims at *learning* (i.e. via optimization) the optimal linear combination of μ weights.

The idea of MKL is summarized in Fig. 10.1: we have $\mathcal{V} = [1, \ldots, m, \ldots, M]$ views of the same data (M blocks of features), which can be spectral bands, groups of bands, image time sequences, or spatial filters of different scale or nature [45, 52]. For each view, we build a separate kernel indexed by m, each one of the most appropriate type and with the most appropriate parameters. We aim at finding the best combination of the form:

$$\begin{aligned} K(\mathbf{x}_i, \mathbf{x}_j) &= \sum_{m=1}^{M} \mu_m K_m(\mathbf{x}_i, \mathbf{x}_j), \\ s.t. \quad & \mu_m \geq 0 \\ & \sum_{m=1}^{M} \mu_m = 1. \end{aligned} \tag{10.24}$$

MKL aims at optimizing a convex linear combination of kernels, i.e., the μ_m weights, at the same time as it trains the classifier. In the case of the SVM introduced in Sect. 10.2.7, the optimization of the μ_m weights involves gradient descent over the SVM objective value [58]. Globally, we adopt a minimization strategy aternating

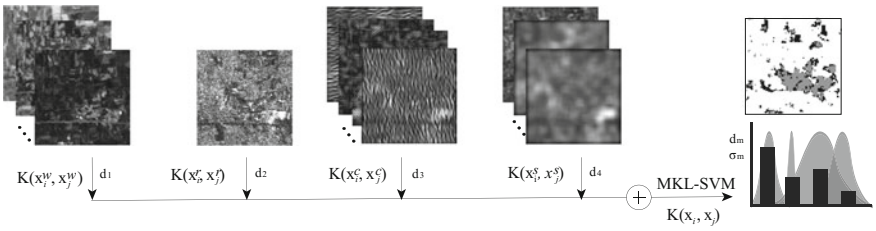


Fig. 10.1 General idea behind multiple kernel learning. Given different sources of registered data, a linear combination of the different similarity matrices (the kernels) is found. (adapted from [45])

two steps: first, we solve an SVM with the composite kernel defined by current μ_m and then we update μ_m by gradient descent.

If we use the kernel in Eq. (10.25), and then plug it into the SVM dual formulation, we obtain the following problem:

$$\max_{\boldsymbol{\alpha}} \left\{ \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \sum_{m=1}^{M} \mu_m K_m(\mathbf{x}_i, \mathbf{x}_j) \right\} \tag{10.25}$$

constrained to $0 \leq \alpha_i \leq C, \sum_i \alpha_i y_i = 0, \forall i = [1, \ldots, n], \sum_m \mu_m = 1$ and $\mu_m \geq 0$. The dual corresponds to a standard SVM in which the kernel is composed of a linear combination of sub-kernels as in Eq. (10.25). One can show (see [58] for details) that maximizing the dual problem in (10.25) is equivalent to solving the problem:

$$\min_{\boldsymbol{\mu}} J(\boldsymbol{\mu}) \quad \text{such that} \quad \sum_{m=1}^{M} \mu_m = 1, \ \mu_m \geq 0 \tag{10.26}$$

where

$$J(\boldsymbol{\mu}) = \begin{cases} \min_{\mathbf{w},b,\xi} \ \frac{1}{2} \sum_{m=1}^{M} \frac{1}{\mu_m} \|\mathbf{w}_m\|^2 + C \sum_{i=1}^{n} \xi_i \\ \text{s.t.} \qquad y_i \left(\sum_{m=1}^{M} \langle \mathbf{w}_m, \boldsymbol{\Phi}_m(\mathbf{x}_i) \rangle + b \right) \geq 1 - \xi_i \\ \qquad\quad \xi_i \geq 0 \end{cases} \tag{10.27}$$

and $\mathbf{w}_m$ represents the weights of the partial decision function of the subproblem m with associated kernel mapping $\boldsymbol{\Phi}_m(\mathbf{x}_i)$. In other words, we have now an objective to optimize by gradient descent over the vector of possible μ_m values. We therefore alternate the solution of an SVM (providing the current error) and the optimization of μ_m. The ℓ_1 norm constraint on the kernel weights forces some d_m coefficients to be zero, thus encouraging sparsity of the solution and natural feature selection.

10.3.1.2 Multitemporal Image Classification with Multiple Kernels

Here, we assess the effectiveness of multi-temporal image classification using SVM and MKL. Potentially interesting information could be obtained from the MKL model related to the temporal relevance of each image and of its features.

In this experiment, we study the potential of MKL for multitemporal image classification with synthetic data. For the generation of realistic synthetic hyperspectral images, we used data from the Compact High Resolution Imaging Spectrometer (CHRIS), which is mounted on board the small satellite platform PROBA (PRoject for On Board Autonomy). The CHRIS sensor provides hyperspectral images in the spectral range from 400 nm to 1050 nm (62 spectral channels for acquisition Mode 1) [59]. The selected image was acquired in the AgriSAR 2006 campaign over the Dem-

min site (Germany) [60]. This image was selected for the study in order to take into account different surface types, patterns, and spatial textures.

From the original image, we generated a time series of 5 synthetic labeled hyperspectral images of size 200 × 200 pixels containing four classes ('forest', "rural urban area", "winter crops", and "summer crops") that vary along time. Details of data simulation can be found in [14]. Two kinds of changes in the spectral signature were simulated: (i) natural spectral variability of the class accounted by the covariance matrix and the random generation of the samples for the different dates and (ii) changes of the class distributions between dates (e.g., due to illumination or atmospheric effects) simulated with a multiplicative factor over the distribution parameters ($\boldsymbol{\mu}_t = \delta_t \boldsymbol{\mu}$ and $\boldsymbol{\Sigma}_t = \delta_t^2 \boldsymbol{\Sigma}$, where $\delta_t = 0.01t + 0.94$, $t = [1, \ldots, 4]$ for all classes). This way, we simulate the situation where a series of images taken in a short time period is available (for instance at different times of the year) and we integrate the different images to avoid problems related to noise, illumination changes and atmospheric effects.

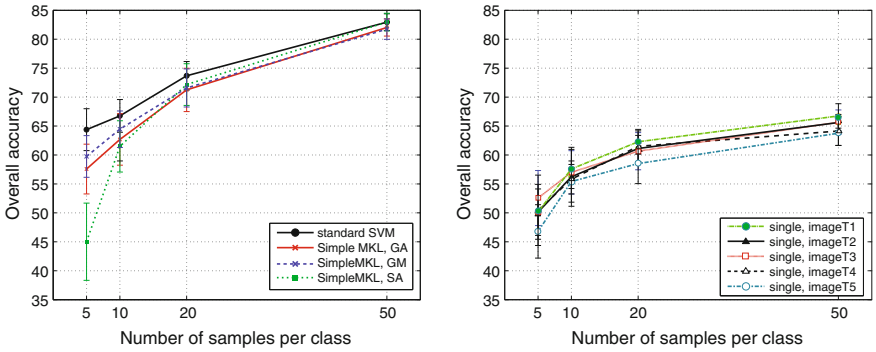
Three experiments using MKL have been run on this data set:

- *GA* (grouped features + model selection by kernel alignment): we use 5 kernels, each one encoding similarity for one of the 5 images of the simulated data. Therefore, each kernel is built on 62 features. Kernel parameters are estimated as those maximizing the kernel alignment [61] between the resulting kernel, $\mathbf{K}$, and the ideal kernel, $\mathbf{y}\mathbf{y}^\top$, that contains values of 1 among samples in $\mathbf{y}$ of the same class and 0 otherwise (normalized using Eq. 10.3):

$$A = \frac{\langle \mathbf{K}, \mathbf{y}\mathbf{y}^\top \rangle_F}{\sqrt{\langle \mathbf{K}, \mathbf{K} \rangle_F \langle \mathbf{y}\mathbf{y}^\top, \mathbf{y}\mathbf{y}^\top \rangle_F}}. \tag{10.28}$$

  In the equation, $\langle \cdot, \cdot \rangle_F$ stands for the Frobenius dot product between matrices, that is, $\langle \mathbf{M}, \mathbf{N} \rangle_F = \sum_{i,j} m_{ij} n_{ij}$. The RBF kernel bandwidth σ_i that maximizes Eq. (10.28) leads to the kernel for group i maximally correlated with the output vector.
- *GM* (grouped features + model selection as in [58]): in this experiment we use four kernels per image.: Instead of optimizing each σ_i using Eq. 10.28, we consider four values of the bandwidth for each group and use each value to build a separate kernel. This leads to 20 kernels (four bandwidths and five images) and therefore to a 20-dimensional weights vectors $\boldsymbol{\mu}$.
- *SA* (single features + model selection by kernel alignment): each one of the 62 × 5 = 310 features is encoded into a separate kernel, with bandwidth σ_i optimized by kernel alignment, as described for the *GA* experiment above.

In all the experiments, the features from the five images have been used to predict the classes of common labels. All the images share the same ground truth. MKL is compared against standard SVMs: On the one hand single models using a single image with a single kernel and, on the other hand, a SVM where all the 310 features have been used into a single kernel. In order to analyze the performance of the proposed



(a) Multitemaporal (b) Single time steps

Fig. 10.2 Overall accuracy curves for the experiments considered

methods under realistic ill-posed situations, we varied the number of training samples *per* class ($n = [5, 10, 15, 20, 30, 40, 50]$) and measured the overall accuracy and the estimated kappa statistic on an independent test set of 15′374 pixels.

Figure 10.2 illustrates the numerical results: the first striking observation is that all single models (shown in Fig. 10.2b) provide unsatisfactory results in terms of accuracy. The performance's gap with the models built using the five images (shown in Fig. 10.2a) increases when the number of labeled samples grows, showing that accounting for several sources of information improves the classification performance. The standard SVM is the model resulting in the highest accuracy for this example, outperforming the MKL models by 4–5% in the experiments using 5 pixels *per* class. But when the number of training pixels grows, the gap is reduced and, from 20 pixels *per* class, the difference becomes numerically insignificant: MKL seems to need a higher number of pixels to converge to a stable solution. The McNemar's tests confirm this hypothesis: When using 5 pixels *per* class, SVM always outperforms MKL ($\bar{z} = 38.78$ for all experiments), but when increasing the number of training pixels to 10, $\bar{z}$ falls to 13.7 and in two experiments over ten, the maps are not significantly different. ($|\bar{z}| < 1.96$). The trend continues with 20 pixels *per* class, where $\bar{z} = 4.5$, a result is statistically the same and MKL outperforms SVM in two of the runs. Using 50 pixels *per* class, the SA experiment even results in the best mean accuracy of all the experiments, but $\bar{z} = 0.01$, showing identical solutions on the average (specifically, 3 solutions are identical, 3 times SVM outperforms MKL, and 4 times MKL results in the best solution).

The classification maps shown in Fig. 10.3 confirm these hypotheses: The use of one image only (Fig. 10.3a, b) results into a noisy classification map and the increase of training pixels does not solve the problem of the poor quality of input information. On the contrary, multitemporal approaches show a strong increase of the quality of the classification maps for both standard SVM (Fig. 10.3c, d) and MKL (Fig. 10.3e, f). Moreover, the higher noise level that can be seen in the MKL solution using 10 pixels *per* class (Fig. 10.3e) is removed when we increase the number of

training examples (Fig. 10.3f). This behavior may be related to the better estimation of the kernel alignment when using a greater number of training pixels.

Even if they perform similarly numerically, MKL has the advantage over SVM of the interpretability of the features importance. This is achieved via the analysis of the $\boldsymbol{\mu}$ weights, which shows the most important images among the five images used (GA experiment) or the features that mostly contribute to the solution (SA experiment).

Looking at the results reported in Fig. 10.4a, all the images participate equally to the solution, maybe with the exception of the image at t_3: on the average, no image is discarded from the final solution. This behavior was expected, because spectral variability has been added to all the images, so that all of them contain a part of useful information for the final solution.

Regarding single features' ranking (Fig. 10.4b), three main groups of features emerge around bands #3, 33, and 53 (centered at 452, 722 and 895 nm, respectively). Interestingly, the same features are selected for the different images (in the barplot of Fig. 10.4b each color corresponds to one of the 5 images): The most important bands (#33, 53) are around the red edge and the NIR zones of the spectrum, and account for the vegetation cell structure, thus helping in characterizing the crops and forest. Bands around #3 account for the blue (tone) of images thus helping in classifying different tonalities in the scene. An important secondary cluster in the band ranking is observed in the red region of the spectrum (between bands #20–30), and around band #60. The first group essentially accounts for the chlorophyll content with red-edge bands. The fact that these bands show relatively lower relevance may be due to the fact that chlorophyll contribution saturates and is transferred to red-edge channels. The secondary subgroup (around #60) characterizes the water content and can be very useful to discriminate between rural urban areas and natural vegetation.

10.3.2 Making Image Representations More Similar with Manifold Alignment

A typical problem in remote sensing classification is to make the algorithms robust to acquisition conditions (e.g., illumination, seasonality, atmosphere), since it is not always possible to run proper atmospheric correction [62, 63] on the data (mainly by the lack of precise ancillary data) and improper correction can harm the results more than improve them. Moreover, atmospheric correction does not solve all problems of discrepancies between acquisitions, since, for instance, local angular effects (BRDF) and seasonal phenological changes still remain and are unknown by a classifier trained on another acquisition (see also Chap. 3). Synergies between physical correction and data driven approaches have proven to be very effective [64].

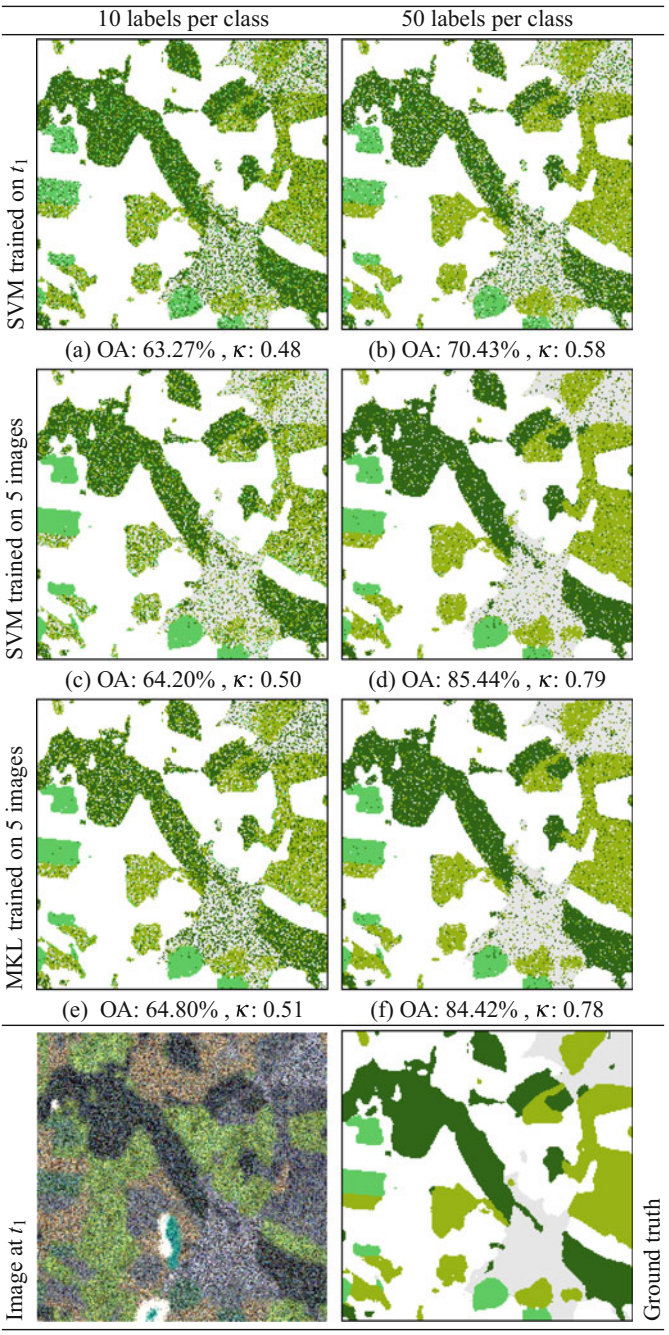


Fig. 10.3 Results for the multitemporal analysis. **a**, **b** Using only the image at t_1, **c**, **d** using all the 5 images into a standard SVM and **e**, **f** using MKL with the SA approach. The *left* column reports results obtained using 10 pixels *per* class, while the *right* column reports results using 50

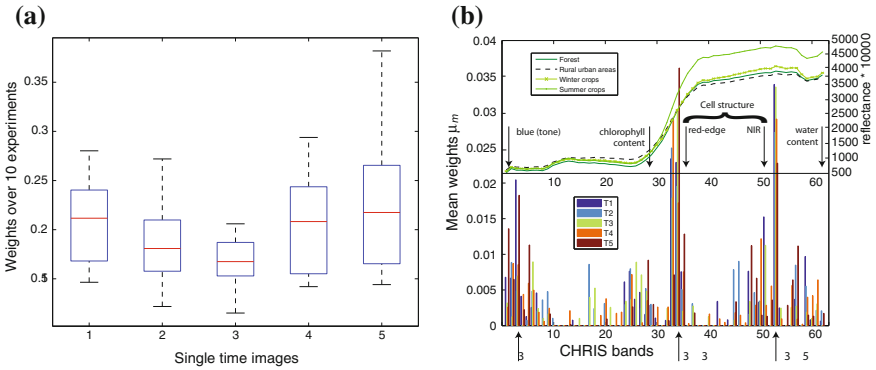


Fig. 10.4 Average μ_m weights (over the 10 experiments) using 50 pixels *per* class for the **a** GA and **b** SA experiments. For the SA experiment, average class spectra are reported for the four classes considered (*right* hand side legend)

When physics-based correction is not possible, one must resort to relative normalization methods, which are data driven and end up providing a matching between the data distributions, but also lose the original physical meaning of the data.

Traditional means for relative-data normalization are histogram matching [65, 66], graph matching [67], and projection methods such as canonical correlation analysis [68].

A second challenging problem in multi-temporal classification is to make classifiers portable across sensors (i.e., be able to reuse a classifier trained on spectra from one sensor to process images acquired by another similar sensor). Such a flexibility is required for many applications, but becomes a necessity when dealing with post-catastrophe intervention, where one cannot wait for the next cloud free acquisition of a specific sensor to run the analysis. A family of methods allowing for such flexibility is again the projective methods based on canonical correlation: methods such as the canonical correlation analysis [68] or the kernel-based canonical correlation analysis [69] allow to align spaces of different dimensionality, because they define a common space where the projections are mostly correlated with the data in their original spaces. These methods never compare the data spaces directly, only the correlations of similarities across the modalities: For this reason, they do not need a true multi-modal metric, which is difficult to obtain in practice.

If CCA-like methods seem to be the answer to the multi-temporal/multi-source problem, they come with a restriction, which is the need of co-registration of all the images to be aligned: (k)CCA are based on the assumption that, at least when defining the projections, each data sample is represented in all the modalities, i.e., each pixel is present and identifiable in all the images. These methods cannot be

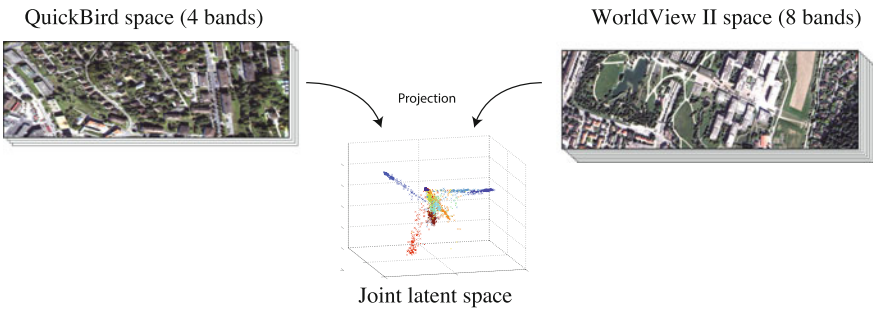


Fig. 10.5 Principle of manifold alignment. Two images from different sensors acquired over different areas are aligned spectrally in a common latent space (in the middle) that maximizes discrimination

used when this requirement is not fulfilled; for example, when considering images of spatially disconnected areas or when dealing with strong geometrical distortions.

Recently, manifold alignment [70] has been proposed as a valid alternative to these methods: Instead of using geographical correspondences across pixels to define the projections, the semi-supervised manifold alignment (SSMA [71]) registers the spectral spaces by using spatially loose registration points: labels. This comes at the price that some labels are required in every domain. In [72], SSMA was successfully used to align sequences of multi-temporal and multi-source (QuickBird and WorldView2) images (Fig. 10.5). Recently, manifold alignment-based methods have known a great success in remote sensing, including methods to account for spatial information [73], to include global and local alignment for classification [74], for correction of illumination effects in hyperspectral images [75] or also for visualization [76]. Despite their success, these methods all work with linear projection functions. The first attempt to use kernel methods with this type of methods is found in [77] and is described below.

10.3.2.1 From Linear to Kernel Manifold Alignment

In this section, we present the KErnel Manifold Alignment method (KEMA) [77]. To explain it in detail, we first review the SSMA method [71]. In the following, we consider a set of M domains, each one of dimensionality d_m (possibly different from one domain to the other), to be aligned. SSMA aligns data from all M domains by projecting them into a common *latent space*. The latent space has two properties: it is discriminant for classification and it respects the original geometry of each manifold (each image in our case). To do so, SSMA minimizes a cost function with three terms: (1) a geometry-preservation term, `GEO`, forcing the local geometry of each manifold to remain unchanged; (2) a similarity term, `SIM`, mapping samples of the same class close to each other; and (3) a dissimilarity term, `DIS`, projecting pixels of different classes far from each other. Loosely speaking, the cost function aims to

maximize $\mathscr{L} = \big((1-\mu)\texttt{SIM}+\mu\texttt{GEO}\big)/\texttt{DIS}$, which involves extracting the smallest eigenvalues from the following generalized eigenproblem [71]:

$$\mathbf{X}(\mathbf{L}_{\texttt{GEO}} + \mu\mathbf{L}_{\texttt{SIM}})\mathbf{X}^\top\boldsymbol{\varphi} = \lambda\mathbf{X}\mathbf{L}_{\texttt{DIS}}\mathbf{X}^\top\boldsymbol{\varphi}, \tag{10.29}$$

where $\boldsymbol{\varphi}$ is the researched common projection matrix of size $d \times d$, with $d = \sum_m d_m$. The rows of $\boldsymbol{\varphi}$ contain a block of projectors for each domain, scaled by $\sqrt{\lambda}$. The matrix $\mathbf{X}$ is a $(d \times n)$ block-diagonal matrix containing the data from the different domains to be aligned, with $n = \sum_m n_m$ being the number of pixels from all domains available to define the projections.

In Eq. (10.29), $\mathbf{L}_{\texttt{GEO}}$, $\mathbf{L}_{\texttt{SIM}}$ and $\mathbf{L}_{\texttt{DIS}}$ are graph Laplacians issued from the similarity matrices $\mathbf{G}$, $\mathbf{S}$ and $\mathbf{D}$, corresponding to the GEO, SIM and DIS terms, respectively [72]. The $\mathbf{G}$ matrix is a block-diagonal matrix containing values 1 if two pixels of a same domain are neighbours in the spectral space and 0 otherwise. Neighbourhood is defined by a proximity graph considering pixels of the same domain only (we do not want to preserve the geometry across domains, only within each domain). The $\mathbf{S}$ matrix summarizes class similarity and contains the value 1 if the pixels are of the same class and 0 otherwise (across all domains). The dissimilarity matrix $\mathbf{D}$ assigns the value 1 if two pixels belong to different classes and 0 otherwise (again, across all domains).

Now, let us kernelize SSMA into KEMA: in the multi-domain setting considered here, we would have to map the M data sets to M Hilbert spaces $\mathscr{H}_m$ of dimension H_m, $\boldsymbol{\phi}_m : \mathbf{x} \mapsto \boldsymbol{\phi}_m(\mathbf{x}) \in \mathscr{H}_m$, $m = [1, \ldots, M]$. The derivation of the kernelized SSMA is presented here with regard to the case $H_m < \infty$ (i.e. $m = [1, \ldots, M]$). Then, we replace all the samples with their mapped feature vectors, to obtain:

$$\boldsymbol{\Phi}(\mathbf{L}_{\texttt{GEO}} + \mu\mathbf{L}_{\texttt{SIM}})\boldsymbol{\Phi}^\top\mathbf{U} = \lambda\boldsymbol{\Phi}\mathbf{L}_{\texttt{DIS}}\boldsymbol{\Phi}^\top\mathbf{U},$$

where $\boldsymbol{\Phi}$ is a block-diagonal matrix containing the data matrices $\boldsymbol{\Phi}_m = [\boldsymbol{\phi}_m(\mathbf{x}_1), \ldots, \boldsymbol{\phi}_m(\mathbf{x}_{n_m})]^\top$ and $\mathbf{U}$ contains the eigenvectors organized in rows for the particular domain defined in Hilbert space $\mathscr{H}_m$, $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_H]^\top$ where $H = \sum_{m=1}^{M} H_m$. Note that $\boldsymbol{\Phi}$ and $\mathbf{U}$ live in a high dimensional space that might be very costly or even impossible to compute. Therefore, we express the eigenvectors as a linear combination of mapped samples using the Representer's theorem [66], $\mathbf{u}_m = \boldsymbol{\Phi}_m\boldsymbol{\alpha}_m$ (or $\mathbf{U} = \boldsymbol{\Phi}\boldsymbol{\Lambda}$ in matrix notation):

$$\mathbf{K}(\mathbf{L}_{\texttt{GEO}} + \mu\mathbf{L}_{\texttt{SIM}})\mathbf{K}\boldsymbol{\Lambda} = \lambda\mathbf{K}\mathbf{L}_{\texttt{DIS}}\mathbf{K}\boldsymbol{\Lambda}, \tag{10.30}$$

where $\mathbf{K}$ is a block-diagonal matrix containing the kernel matrices $\mathbf{K}_m$. Now the eigenproblem becomes of size $n \times n$ instead of $d \times d$, and we can extract a maximum of n components.

This dual formulation is advantageous when dealing with very high dimensional data sets, $d \gg n$, for which the SSMA problem is not well-conditioned. Projection to the latent space requires first mapping the data into the latent space by computing

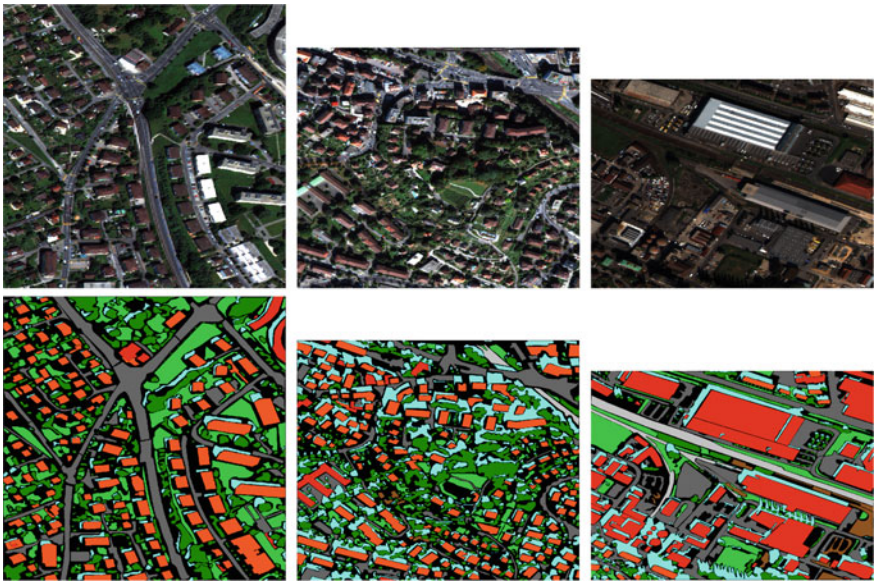


Fig. 10.6 The WorldView-2 images used in the remote sensing image classification experiments. **Color legend:** residential, meadows, trees, roads, shadows, commercial building, railway, bare soil

the corresponding kernel $\mathbf{K}_{im}$ and then applying the projection vector $\boldsymbol{\alpha}_m$ defined therein:

$$\mathbf{x}_i^* = \mathbf{u}_m^\top \boldsymbol{\Phi}_{im} = \boldsymbol{\alpha}_m^\top \boldsymbol{\Phi}_m^\top \boldsymbol{\Phi}_{im} = \boldsymbol{\alpha}_m^\top \mathbf{K}_{im}, \tag{10.31}$$

where $\mathbf{K}_{im}$ is a vector of kernel evaluations between sample $\mathbf{x}_i$ and all samples from domain m used to define the projections $\boldsymbol{\alpha}_m$.

10.3.2.2 Multi-temporal Transfer of Classifiers at Very High Resolution

In this application, we consider the problem of classifying pixels from a series of multispectral remote sensing images into a limited number of land cover classes. The images in this experiment are three subsets taken from two WorldView-2 scenes with spatial resolution of 2.4 m and eight spectral channels in RGB and infrared wavelengths. All regions imaged are located in the city of Lausanne, Switzerland: the Montelly and Malley images are subsets of images acquired the September 29, 2010, while the Prilly subset is part of a scene acquired the August 2, 2011. All scenes have been pan sharpened using the Gram-Schmidt transform to reach a 0.6 m resolution. Figure 10.6 illustrates the RGB composites and the exhaustive ground truth of these data sets.

We consider the following standard domain adaptation scenario: We assume that one of the images has sufficient label information and we consider it in turn as the leading domain with $l_1 = 100$ labelled pixels per class. The objective is to have a classifier that can predict well on all three images with minimal effort in terrestrial campaigns or photo-interpretation on the new acquisitions, i.e., the least number of new labelled pixels from these (new) incoming domains. Therefore, the two other domains provide less labelled pixels, with $l_2 = l_3 = [10, \ldots, 90]$. Each domain also provides 500 unlabeled pixels. All domains are aligned together with either SSMA or KEMA (with a RBF kernel with the parameter σ_m set as half of the median Euclidean distance between all pixels in domain m) and then a linear SVM is trained on the aligned domain using the same labelled pixels used to define the projections. The SVM regularization parameter was tuned by tenfold cross-validation on the training set.

Table 10.1 shows the results. We compared KEMA with the following competing approaches: the case where no adaptation is performed ('Raw'), the SSMA algorithm, and the case where only labelled pixels from the domain to be predicted are used to train the linear SVM ("Test"). In the table, each row block corresponds to an image used as the leading one for training and each column to the domain used for validation.

First, the case without adaptation is related to a sharp loss in accuracy (up to 30% with respect to the "Test" case), which shows the data set shift in the data and the need for adaptation strategies. The results observed in the rest of the table confirm that KEMA improves the results of SSMA in almost all cases, with an increase of up to +4% in accuracy with respect to SSMA and of up to +8% with respect to the case without adaptation. As for SSMA, KEMA always outperforms the results without adaptation and is insensitive to the presence of labelled samples from other domains, when predicting in the source domain itself. Note that the "Test" baseline is often surpassed, indicating the ability to extract discriminative (SSMA, KEMA) and nonlinear (KEMA) features allowing for precise pixelwise land-cover categorization. An extension of this study to the alignment multi-source images can be found in [75] and also shows the potential of KEMA as a nonlinear feature extractor aligning heterogeneous domains.

10.3.3 New Challenges

In this section, we reviewed two frameworks to perform multi-modal image classification: the multiple kernel learning (MKL), where a kernel function over multiple data sources is learned as a linear combination of kernels specializing over single sources, and the manifold alignment, where data spaces are made more similar by projecting them into a common latent space. Of course, the new cutting edge research areas in kernel methods for classification exceed these two topics.

The inclusion of spatial information has always played a major role in remote sensing image analysis: some examples on how to include such information have been

Table 10.1 Linear SVM classification accuracies for the multitemporal experiment (Mo: Montelly; Pr: Prilly; Ma: Malley (see Fig. 10.6)). SSMA results from [72]

		Test image											
			Mo				Pr				Ma			
		$l_2 = l_3$	0	10	50	90	0	10	50	90	0	10	50	90
Training image	Mo	Raw	71.6	71.5	69.3	68.8	52.9	62.5	72.5	74.1	47.1	50.1	54.6	55.8
		SSMA	–	72.6	72.9	72.6	–	72.6	75.6	76.9	–	54.9	61.2	61.0
		KEMA	–	**74.9**	74.7	74.5	–	69.7	78.1	**79.2**	–	57.4	62.0	**62.9**
	Pr	Raw	58.6	64.2	67.3	68.2	74.1	75.2	75.3	74.4	48.3	54.0	56.0	55.9
		SSMA	–	65.7	72.0	72.4	–	76.5	76.6	76.8	–	54.7	61.1	60.9
		KEMA	–	63.3	73.7	**75.0**	–	**79.5**	79.2	79.2	–	52.9	61.7	**63.0**
	Ma	Raw	65.7	66.0	67.6	68.3	49.1	65.0	72.3	74.2	60.3	58.2	56.6	56.0
		SSMA	–	65.0	71.3	72.4		71.8	75.4	76.9	–	61.2	61.9	61.5
		KEMA	–	62.2	72.9	**74.5**	–	65.4	78.2	**79.3**	–	63.5	63.5	**63.7**
		Test	71.6	–	–	–	74.1	–	–	–	60.3	–	–	–

presented above, when introducing the multiple kernel learning framework [12], both approaches based on Markov and conditional random fields (cf. Chaps. 4 and 7) can be deployed for enforcing a prediction consisted in the space of outputs (these methods are also known as *structured output* models): in [78, 79], contrast-sensitive conditional random fields models are used to enforce the smoothness on the classification map, where the contrast sensitivity is actually defined by a Gaussian kernel between pixels. In [80], the Markovianity assumption is directly encoded in the kernel.

Another new research avenue pushing the structured outputs logic is to learn the parameters of the structure model: Instead of enforcing smoothness or contrast sensitivity assumptions on the output space, one could *learn* the correlation ranges in the outputs spaces [81], the dependencies between classes (e.g., as a hierarchy [82]) or learn a whole conditional random field model based on classes co-occurrence structure [83]. Models such as the structured SVMs are well suited to the task and are gaining momentum in the remote sensing community.

Subspace learning using kernel methods has also emerged as a new challenger to discriminative models such as the support vector machine: the subspaces where the data live do not have to be linear and kernel methods are a natural way of describing complex manifold geometries (nonlinear, intersecting, and multimanifold) via kernel projection-then-linear description strategies. Classifiers based on collaborative representation [84, 85] and subspace-regularized graphs [86] are only few of the recent examples of advances in kernel methods exploring the approximation of data manifolds via the use of kernel functions.

10.4 Biophysical Parameter Estimation

In this section, we review some recent advances in Gaussian processes regression especially suited for biophysical parameter retrieval from optical remote sensing data. In particular, we will review the main aspects to design covariance functions that capture non-stationarities and multiscale time relations, as well as GPR that can learn signal-to-noise relations, rank features, and derive confidence intervals for the predictions.

10.4.1 Covariances in Gaussian Processes

10.4.1.1 Structured, Non-stationary and Multiscale

Commonly used kernels families include the squared exponential (SE), periodic (Per), linear (Lin), and rational quadratic (RQ) [87]. Figure 10.7 shows the base kernels and drawings from the GP prior. These base kernels can be actually combined following simple operations: summation, multiplication, or convolution (see Sect. 10.2.5). This way one may build sophisticated covariances from simpler ones as

seen before in the MKL framework. Note that the same essential property of kernel methods apply here: a valid covariance function must be positive semidefinite. In general, the design of the kernel should rely on the information that we have for each estimation problem and should be designed to get the most accurate solution with the least amount of samples.

In Fig. 10.7a–d, one-dimensional base kernels are presented. Nevertheless, kernels over multidimensional inputs can be actually constructed by adding and multiplying kernels over individual dimensions. See [87] for the explicit functional form of each kernel. Some simple kernel combinations are also represented in Fig. 10.7: a linear plus periodic covariances may capture structures that are periodic with a trend (Fig. 10.7e), while a linear plus squared exponential covariance can accommodate structures with increasing variation (Fig. 10.7f).

By summing kernels, we can model the data as a superposition of independent functions, possibly representing different structures in the data. For example, in multitemporal image analysis, one could, for instance, dedicate a kernel for the time domain (trying to capture trends and seasonal effects) and another kernel function for the spatial domain (equivalently capturing spatial patterns and auto-correlations).

In time-series models, sums of kernels can express superposition of different processes, possibly operating at different scales: very often changes in geophysical variables through time occur at different temporal resolutions (hours, days, etc.), and this can be incorporated in the prior covariance with those simple operations. In multiple dimensions, summing kernels gives additive structure over different dimensions, similar to generalized additive models [88]. Alternatively, multiplying kernels allows us to account for interactions between different input dimensions or different notions of similarity. In the following section, we show how to design kernels that incorporate particular time resolutions, trends, and periodicities.

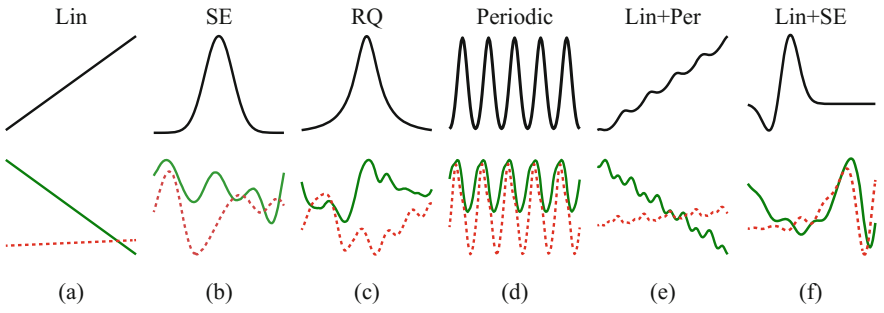


Fig. 10.7 Base kernels (*top*) and two random draws from a GP with each respective kernel (*bottom*) (adapted from [87])

10.4.1.2 Time-Based Covariance for GPR

Signals to be processed typically show particular characteristics, with time-dependent cycles and trends. One could include time t_i as an additional feature in the definition of the input samples. This *stacked approach* [12] (cf. Chaps. 6 and 7) essentially relies on a covariance function $k(\mathbf{z}_i, \mathbf{z}_j)$, where $\mathbf{z}_i = [t_i, \mathbf{x}_i]^\top$. The shortcoming is that the time relations are naively left to the nonlinear regression algorithm, and hence no explicit time structure model is assumed. To cope with this, one can use a linear combination (or composite) of different kernels: one dedicated to capture the different temporal characteristics and the other to the feature-based relations.

The issue here is how to design kernels capable of dealing with non-stationary processes. A possible approach is to use a *stationary* covariance operating on the variable of interest after being mapped with a nonlinear function engineered to discount such undesired variations. This approach was used in [89] to model *spatial patterns* of solar radiation with GPR. It is also possible to adopt a squared exponential (SE) as stationary covariance acting on the *time* variable mapped to a two-dimensional *periodic space* $\mathbf{z}(t) = [\cos(t), \sin(t))]^\top$, as explained in [3]:

$$k(t_i, t_j) = \exp\left(- \frac{\|\mathbf{z}(t_i) - \mathbf{z}(t_j)\|^2}{2\sigma_t^2} \right), \tag{10.32}$$

which gives rise to the following periodic covariance function:

$$k(t_i, t_j) = \exp\left(- \frac{2\sin^2[(t_i - t_j)/2]}{\sigma_t^2} \right), \tag{10.33}$$

where σ_t is a hyperparameter characterizing the periodic scale and needs to be inferred. It is not clear, though, that the seasonal trend is exactly periodic, so we modify this equation by taking the product with a squared exponential component, to allow a decay away from exact periodicity:

$$k_2(t_i, t_j) = \gamma \exp\left(- \frac{2\sin^2[\pi(t_i - t_j)]}{\sigma_t^2} - \frac{(t_i - t_j)^2}{2\sigma_d^2} \right), \tag{10.34}$$

where γ gives the magnitude, σ_t the smoothness of the periodic component, σ_d represents the *decay-time* for the periodic component, and the period has been fixed to one year. Therefore, our final covariance is expressed as:

$$k([\mathbf{x}_i, t_i], [\mathbf{x}_j, t_j]) = k_1(\mathbf{x}_i, \mathbf{x}_j) + k_2(t_i, t_j), \tag{10.35}$$

which is parameterized by only three more hyperparameters collected in $\boldsymbol{\theta}$. Note that this kernel function allows us to incorporate time easily, but the relations between time t_i and signal $\mathbf{x}_i$ samples is missing. Some approximations to deal with this issue exist in the literature, such as cross-kernel composition [12, 90] or latent force models [91].

Table 10.2 Results for the estimation of the daily solar irradiation of linear and nonlinear regression models. Subscript METHOD$_t$ indicates that the METHOD includes time as input variable. Best results are highlighted in bold, the second best in italics

METHOD	ME	RMSE	MAE	R
RLR	0.27	4.42	3.51	0.76
RLR$_t$	0.25	4.33	3.42	0.78
SVR [92]	0.54	4.40	3.35	0.77
SVR$_t$	0.42	4.23	3.12	0.79
RVM [93]	0.19	4.06	3.25	0.80
RVM$_t$	0.14	3.71	3.11	0.81
GPR [3]	0.14	3.22	2.47	*0.88*
GPR$_t$	*0.13*	*3.15*	*2.27*	*0.88*
TGPR	**0.11**	**3.14**	**2.19**	**0.90**

We show the advantage of encoding such prior knowledge and structure in the relevant problem of solar irradiation prediction using GPR. Noting the non-stationary temporal behaviour of the signal, we develop a particular time-based composite covariance to account for the relevant seasonal signal variations. Data from the AEMET radiometric observatory of Murcia (Southern Spain, 38.0° N, 1.2° W) were used. Table 10.2 reports the results obtained with GPR models and several statistical regression methods: regularized linear regression (RLR), support vector regression (SVR), relevance vector machine (RVM), and GPR. All methods were run with and without using two additional dummy time features containing the year and day-of-year (DOY). We will indicate the former case with a subscript t, e.g., SVR$_t$. From the numerical results, we come to three observations. First, including time information improves all baseline models. Second, the best overall results are obtained by the GPR models, when including time information or not. Third, the proposed temporal GPR (TGPR) outperforms the rest (including GPR and GRP$_t$) in all quality measures.

10.4.2 Ranking Features Through the Automatic Relevance Determination (ARD) Covariance

One of the advantages of GPs is that during the development of the GP model, the predictive power of each single band is evaluated for the parameter of interest through calculation of the ARD [3], which is expressed a

$$K(\mathbf{x}_i, \mathbf{x}_j) = \nu \exp\left(-\sum_{m=1}^{d} \frac{(x_i^f - x_j^f)^2}{2\sigma_f^2} \right) + \sigma_n^2 \delta_{ij},$$

where x_i^f represents the feature (band) f of the input vector $\mathbf{x}_i$, ν is a scaling factor, σ_n is the standard deviation of the (estimated) noise, and a σ_f is the length scale per input features, $f = 1, \ldots, d$. This is a very flexible covariance function that

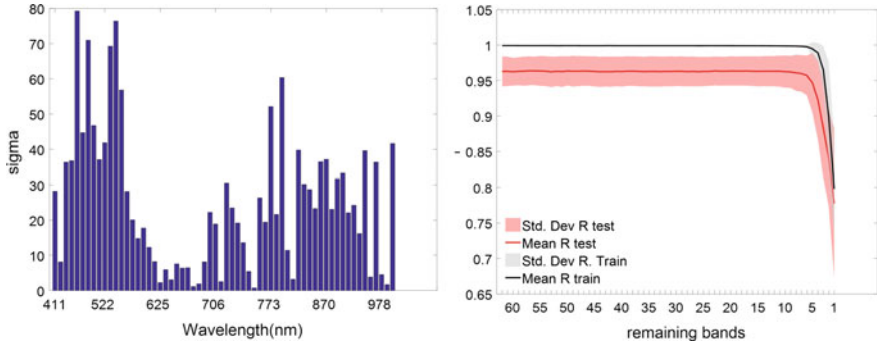


Fig. 10.8 Estimated σ_f values for one GP model using 62 CHRIS bands (*left*). The lower the σ_f the more important the band is for regression. Chl r and standard deviation (SD) of training and validation for GP fittings using sequential backward band removal (*right*) (adapted from [94, 95])

typically suffices to tackle most of the problems. However, note that a SE typically can approximate smoothly varying functions, which may not be the case in particular problems.

Specifically, band ranking through σ_f may reveal the bands that contribute the most to the development of a GP model. An example of the σ_f's for one GP model trained with field leaf chlorophyll content (Chl) data and with 62 CHRIS bands is shown in Fig. 10.8 (left) [94]. The band with highest σ_f is the least contributing to the model. It can be noted that a relatively few bands (about 8) were evaluated as crucial for Chl estimation, while the majority of bands were evaluated as less contributing. The figure also suggests that the most relevant spectral region is to be found between 625 and 1000 nm. Most contributing bands were positioned at the red and the red edge, at 625 and 730 nm respectively, but not all bands within the red edge were evaluated as relevant. This is due to when having a large number of bands available then neighbouring bands do not provide much additional information and can thus be considered as redundant.

This does not necessarily mean that other bands are obstructing optimized accuracies. By applying a simple iterative backward greedy algorithm, in which the impact of the inputs on the prediction error is evaluated in the context or absence of the other predictors, the most informative bands and the least numbers of bands that preserve optimized accuracies are identified. Practically, at each iteration, the least significant band was removed, i.e., the one with the highest σ_f, and a new GPR model was trained with the remaining bands only. This *sequential backward band removal* (SBBR) algorithm is similar to the one often applied in classification using SVM, referred to as *recursive feature elimination* (RFE). In RFE, the feature with the smallest ranking score is eliminated in order to recursively remove insignificant features. However, RFE is merely interested in determining optimized classification results [96–98], while in SBBR we remove backwards until only one band is left. When subsequently plotting goodness-of-fit statistics over the iterations (here r) it can be inspected that results kept stable starting from using all bands until a few

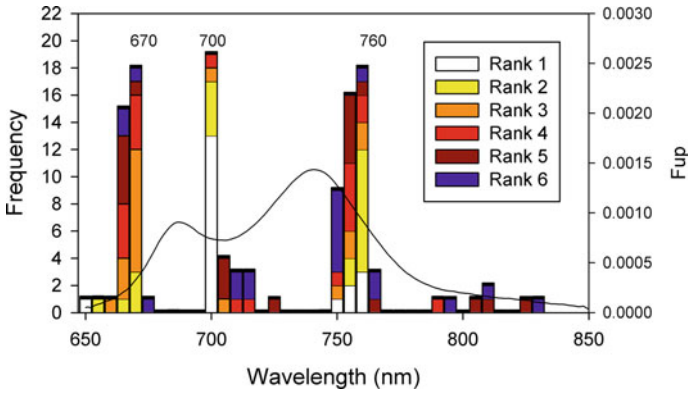


Fig. 10.9 Frequency plots of the top eight ranked bands with lowest σ_f values in 20 runs of GPR prediction of Chl based on upward fluorescence (F_{up}) emission. An emission curve is given as illustration (from [100])

bands were left. Only when less than 4 bands were left accuracies started to degrade rapidly, as observed in Fig. 10.8 (right).

Consequently, the SBBR algorithm proved to be a valuable tool to detect the minimum number of most sensitive bands of a sensor towards a biophysical parameter.

A similar analysis was applied by sorting simulated Sentinel-2 bands on their relevance and counting the band rankings over 50 repetitions. In [32], the four most relevant bands were tracked for Chl, LAI and fCOVER and for different Sentinel-2 settings. It demonstrated the potential of Sentinel-2, with its new band in the red edge, for vegetation properties estimation. Also in [99], σ_f were used to analyze band sensitivity of Sentinel-2 towards LAI. A similar approach was pursued on analyzing leaf Chl based on tracking the most sensitive spectral regions of sun-induced fluorescence data [100], as displayed in Fig. 10.9.

The SBBR algorithm has recently been automated into the GPR band analysis tool (GPR-BAT, Fig. 10.10) [101]. GPR-BAT is implemented within the framework of the free ARTMO's MLRA (machine learning regression algorithms) toolbox [102], which is dedicated to the transformation of optical remote sensing images into biophysical products.[1] The main purpose of GPR-BAT is to identify how many bands are minimally needed in order to retain robust results and what are the most sensitive wavelengths. Although emphasis is on vegetation properties, essentially GPR-BAT can be applied to any measured (or modeled) surface biophysical or geophysical variable when associated with spectral data.

[1] http://ipl.uv.es/artmo/.

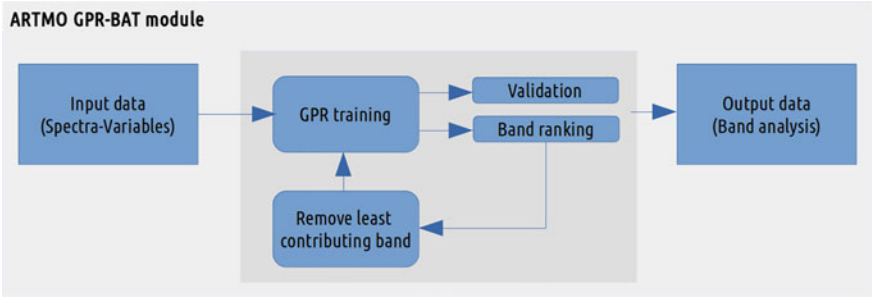
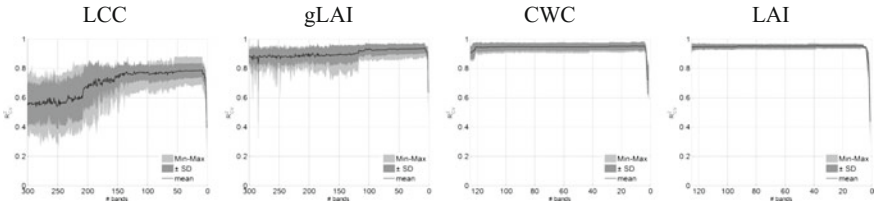


Fig. 10.10 Schematic flow diagram of GPR-BAT within ARTMO's MLRA toolbox (from [101])

To illustrate the utility of GPR-BAT, two hyperspectral data sets were analyzed to highlight the most informative bands: (1) a field hyperspectral data set (400–1100 nm at 2 nm resolution: 301 bands) with leaf chlorophyll content (LCC) and green leaf area index (gLAI) collected for maize and soybean (Nebraska, USA); and (2) an airborne HyMap data set (430–2490 nm: 125 bands) with leaf area index (LAI) and canopy water content (CWC) collected for a variety of crops (Barrax, Spain). For each of these biophysical variables, optimized retrieval accuracies can be achieved with just 4–9 well-identified bands, and performance was improved over using all bands, as displayed in Fig. 10.11.

Especially when many bands are involved, as in case of the field hyperspectral data set, improvements are significant: using all bands performed almost as poor as when using only one band. Generally, entering all hyperspectral bands into a regression model performs poorly due to multi-collinearity and inclusion of noisy bands. To overcome this, either band selection or spectral reduction techniques are recommended.

GPR-BAT results suggest that band redundancy is less an issue when reducing to superspectral data (typically <50 bands), and using all those bands can equally lead to top-performing regression models. Accordingly, from a GPR point of view, for multispectral and superspectral sensors it seems to be enough to have the spectral bands rightly located along the spectral range. Interestingly, for none of the tested variables using only the two best-performing bands led to optimal results. This suggests that applying two-band indices to hyperspectral data is suboptimal in exploiting the embedded information content and is therefore not recommended.



(a) field spectroradiometer (b) airborne imaging spectrometer

Fig. 10.11 Cross-validation R^2_{CV} statistics (mean, standard deviation, and min-max ranges) for **a** field LCC and gLAI and **b** airborne CWC and LAI. The plots show the sequential removal of the least contributing band using GPR-BAT (from [101])

10.4.3 Uncertainty Intervals

In this section, we use GPR models for retrieval and portability in space and time. For this, we will exploit the associated predictive variance (i.e., uncertainty interval) provided by GPR models. Consequently, retrievals with high uncertainties refer to pixel spectral information that deviates from what has been represented during the training phase. In turn, low uncertainties refer to pixels that were well represented in the training phase.

> The quantification of variable-associated uncertainties is a strong requirement when remote sensing products are ingested in higher level processing, e.g., to estimate ecosystem respiration, photosynthetic activity, or carbon sequestration [103].

The application of GPR for the estimation of biophysical parameters was initially demonstrated in [94]. A locally collected field data set called SPARC-2003 at Barrax (Spain) was used for training and validation of GPR for the vegetation parameters of LAI, Chl, and fractional vegeatation cover (fCOVER). Sufficiently high-validation accuracies were obtained ($r^2 > 0.86$) for processing a CHRIS image into these parameters, as shown in Fig. 10.12. Within the uncertainty maps, areas with reliable retrievals are clearly distinguished from areas with unreliable retrievals. Low uncertainties were found on irrigated areas and harvested fields. High uncertainties were found on areas with remarkably different spectra, such as bright, whitish calcareous soils, or harvested fields. This indicates that the input spectrum deviates from what has been presented during the training stage, thereby imposing uncertainties to the retrieval. It also suggests that those non-vegetated areas would benefit from additional sampling in order to make the GPR model more generally applicable.

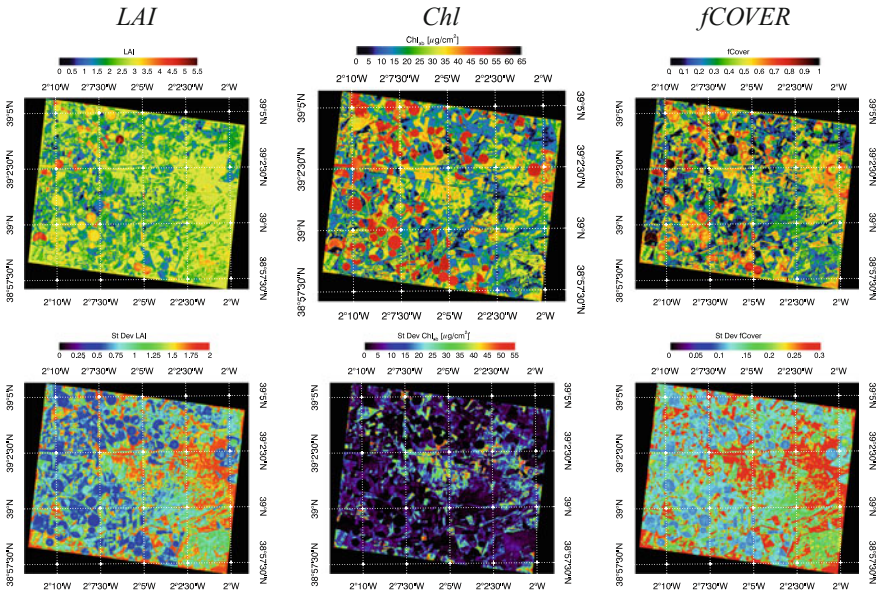


Fig. 10.12 Prediction maps (*top*) and associated uncertainty intervals (*bottom*), generated with GP and four bands of the CHRIS 12-07-2003 nadir image (adapted from [94])

GPR models were subsequently applied to the SPARC data set, but after re-sampling to different Sentinel-2 band settings and uncertainties were inspected [32]. On the whole, adding spectral information from 4 bands at 10 m to 10 bands at 20 m led to reduction of uncertainties and thus more meaningful biophysical parameter maps. The locally trained GP models were applied to simulated Sentinel-2 images in a follow-up study [104]. Time series over the local Barrax site as well images across the world were processed. Also the role of an extended training data set by adding spectra of non-vegetated surfaces were evaluated. Subsequently the uncertainty values were analyzed. By using the extended training data set not only further improved performances but also allowed a decrease in theoretical uncertainties. The GPR models were successfully applied to simulated Sentinel-2 images covering various sites; associated relative uncertainties were on the same order as those generated by the reference image.

The generally low uncertainty intervals over vegetated surfaces suggest that the locally trained GPR models are portable to other sites and conditions.

Remaining large uncertainties within images were due to surface heterogeneity, and associated spectral heterogeneity increased with a finer spatial resolution (i.e., Sentinel-2 provides images with a resolution of 10 m). It appeared that at 10 m reso-

lution, there is a greater problem of portability within an image than to other images with vegetation cover. Hence, from a practical perspective, this implies that uncertainty maps serve as a useful quality layer that allows masking out surfaces that fall beyond an acceptable uncertainty.

As a final example, uncertainty estimates were exploited to assess the robustness of the retrievals at multiple spatial scales. In [95], retrievals from hyperspectral airborne and spaceborne data over the Barrax area were compared. Based on the spareborne SPARC-2003 data set, GPR developed a model that was excellently validated (r^2: 0.96). The SPARC-trained GPR model was subsequently applied to airborne CASI flightlines (Barrax, 2009) to generate Chl maps. The accompanying uncertainty maps provided insight into the robustness of the retrievals. In general similar uncertainties were achieved by both sensors, which is encouraging for upscaling estimates from field to landscape scale. The high-spatial resolution of CASI in combination with the uncertainties allows us to observe the spatial patterns of retrievals in more detail. Some examples of mean estimates and associated uncertainties are shown in Fig. 10.13. The GPR uncertainty maps immediately highlighted the areas where the user should be careful with the obtained Chl estimates, such as over man-made surfaces, due to irrigation or variations in soil properties. It is expected that in the near future this extra source of information freely offered by GPR will become increasingly important when GPR-delivered vegetation properties estimates will become operationally available, e.g., into a smartphone app [105].

10.4.4 New Challenges with GPR

Research on GPR is still very active and challenging new questions remain. A new research avenue using GPR models involves the development of meta-models also named *emulators*. Emulation is a technique used to estimate simulations outcomes when the computer model under investigation is too computationally costly to be run a sufficient number of times [106]. Emulators approximate the functioning of deterministic (physical) models through statistical learning regression methods. Because of their computational efficiency and outstanding accuracy, GPRs have been a first choice in developing emulators [106–108]. Although the concept of emulators is known within statistical and computer sciences [109–111], their use is only at its infancy in optical remote sensing. Recent pioneer studies [112, 113] showed that emulators can successfully approximate physical vegetation and atmosphere radiative transfer models (RTMs). RTM emulation functions essentially the same as retrieval through machine learning regression, but instead of delivering biophysical variables as outputs they are used as input in the regression model, and spectral data is generated as output. Probably the most significant advantage of emulators is the tremendous gain in processing speed, in the orders of tens to ten thousands depending on the speed of the original RTM. At the same time, they hardly require memory space, since only a few model coefficients are stored for prediction. Consequently, emulators can become an attractive technique for a diversity of remote sensing appli-

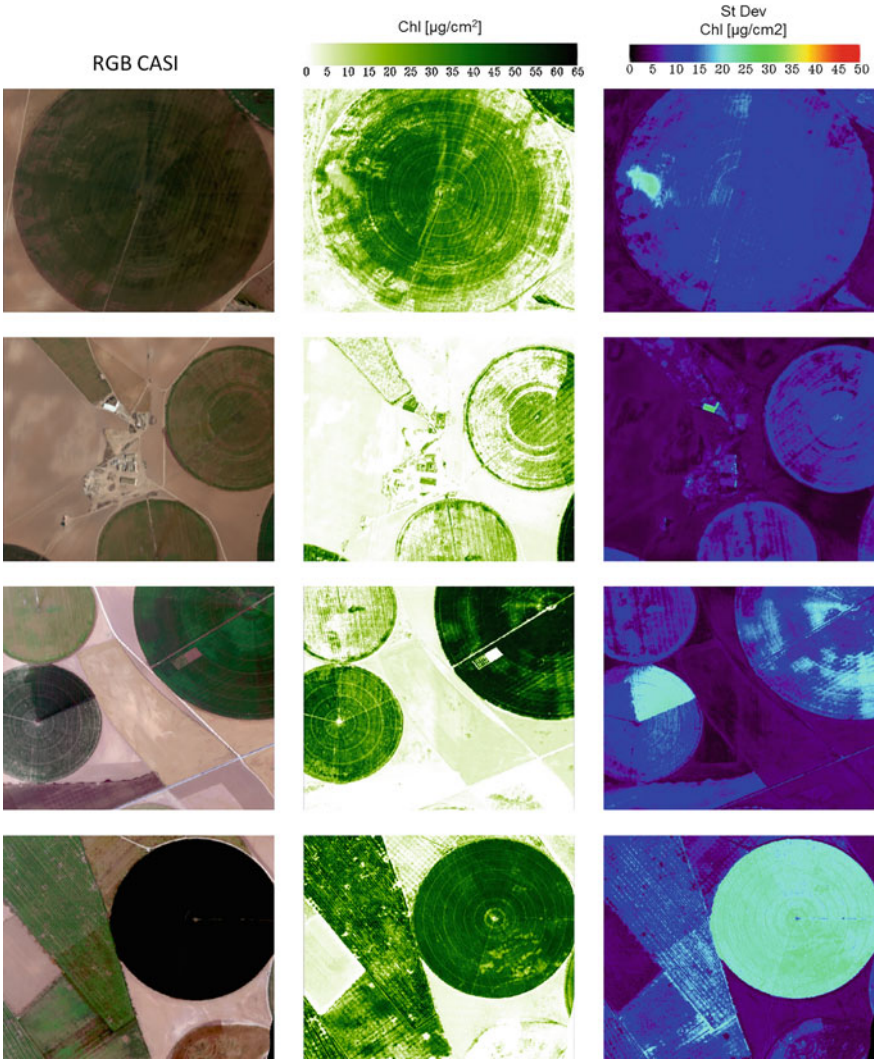


Fig. 10.13 Four examples of CASI RGB snapshots [*left*], *Chl* estimates [*middle*], and related uncertainty intervals [*right*] (adapted from [95])

cations, for instance applying global sensitivity analysis (GSA) on computationally expensive RTMs such as Monte Carlo ray tracing models. GSA typically requires many thousand simulations, which caused that so far only computationally cheap RTMs were evaluated. In turn, GSA can be applied to an RTM-like emulator and deliver results quasi-instantly [114].

Emulators can also become highly beneficial with respect to biophysical parameter retrieval through inversion of physical models. RTMs are traditionally used in

inversion schemes against optical images through iterative optimization or through look-up tables (LUTs). The optimization consists in minimizing a cost function, which estimates the difference between measured and estimated variables by successive input variable iteration. Such iterative optimization algorithms are computationally demanding on a pixel-by-pixel basis, and hence time-consuming when large remotely sensed images are inverted. Accordingly, when replacing the original RTMs by their emulated counterparts, numerical inversion schemes can again become an efficient alternative. It would bypass the need to invest in large and heavy LUTs and instead open opportunities to include emulators of advanced, computationally expensive RTMs into retrieval schemes. To conclude with, emulation emerge as a promising method to fully aboard the problem of developing flexible statistical models that discover and incorporate physical knowledge about the problem. We expect more exciting developments in that area of intersection between physics and machine intelligence.

10.5 Conclusions

In this chapter, we provided an introduction to kernel methods for remote sensing image analysis and presented some new developments in this area of research. We focused on the challenging problems of multimodal semantic image labeling (i.e., the act of assigning each pixel to a semantic (land cover) class by using a variety of data sources) and retrieval of biophysical parameters using different Gaussian processes regression models accounting for structures and prior information about the data. For both classification and regression, we presented recent research efforts, underlying the role of kernel methods in the improvement of the performances and interpretability of the modelling stage by explicitly modelling data and problem specificities.

We believe that such inclusion of domain knowledge, along with the ability to reuse existing field data across several images acquisition is key to the success of modern remote sensing data analysis tasks, where analysts are confronted with data coming at high speed and volume, but also from heterogeneous sources.

Acknowledgements The authors wish to deeply acknowledge the collaboration, comments, and fruitful discussions with many researchers during the last decade on GP models for remote sensing and geoscience applications: Miguel Lázaro-Gredilla (Vicarious), Robert Jenssen (Univ. Tromsø, Norway), Martin Jung (MPI, Jena, Germany), and Sancho Salcedo-Saez (Univ. Alcalá, Madrid, Spain).

This work has been partly supported by the Swiss National Science Foundation (grant PZ00P2-136827, http://p3.snf.ch/project-136827), by the Spanish Ministry of Economy and Competitiveness under project ESP2013-48458-C4-1-P, and the European Research Council (ERC) under the ERC-CoG-2014 SEDAL under grant agreement 647423.

References

1. Schölkopf, B., Smola, A.J.: Learning with Kernels. MIT Press, Cambridge (2002)
2. Boser, B.E., Guyon, I.M., Vapnik, V.N.: A training algorithm for optimal margin classifiers. In: Haussler, D. (ed.) 5th Annual ACM Workshop on COLT, pp. 144–152. ACM Press, Pittsburgh, PA (1992)
3. Rasmussen, C.E., Williams, C.K.I.: Gaussian Processes for Mach Learn. The MIT Press, New York (2006)
4. Momma, M., Bennet, K.: Sparse kernel partial least squares regression. In: Proceedings of Conference on Learning Theory, COLT (2003)
5. Belkin, M., Niyogi, P., Sindhwani, V.: Manifold regularization: a geometric framework for learning from labeled and unlabeled examples. J. Mach. Learn. Res. **7**, 2399–2434 (2006)
6. Camps-Valls, G., Bruzzone, L.: Kernel-based methods for hyperspectral image classification. IEEE Trans. Geosci. Remote. Sens. **43**, 1351–1362 (2005)
7. Camps-Valls, G.: New machine-learning paradigm provides advantages for remote sensing. SPIE Newsroom (2008)
8. Melgani, F., Bruzzone, L.: Classification of hyperspectral remote sensing images with support vector machines. IEEE Trans. Geosci. Remote. Sens. **42**, 1778–1790 (2004)
9. Waske, B., Benediktsson, J.A.: Fusion of support vector machines for classification of multisensor data. IEEE Trans. Geosci. Remote. Sens. **45**, 3858–3866 (2007)
10. Foody, G.M., Mathur, A.: Toward intelligent training of supervised image classifications: directing training data acquisition for SVM classification. Remote. Sens. Environ. **93**, 107–117 (2004)
11. Chi, M., Feng, R., Bruzzone, L.: Classification of hyperspectral remote-sensing data with primal SVM for small-sized training dataset problem. Adv. Space Res. **41**(11), 1793–1799 (2008)
12. Camps-Valls, G., Gómez-Chova, L., Muñoz-Marí, J., Vila-Francés, J., Calpe-Maravilla, J.: Composite kernels for hyperspectral image classification. IEEE Geosci. Remote. Sens. Lett. **3**, 93–97 (2006)
13. Tuia, D., Ratle, F., Pozdnoukhov, A., Camps-Valls, G.: Multi-source composite kernels for urban image classification. IEEE Geosci. Remote. Sens. Lett. **7**, 88–92 (2010)
14. Camps-Valls, G., Gómez-Chova, L., Muñoz-Marí, J., Rojo-Álvarez, J., Martínez-Ramón, M.: Kernel-based framework for multitemporal and multisource remote sensing data classification and change detection. IEEE Trans. Geosci. Remote. Sens. **46**, 1822–1835 (2008). cited By 148
15. Plaza, A., Benediktsson, J.A., Boardman, J., Brazile, J., Bruzzone, L., Camps-Valls, G., Chanussot, J., Fauvel, M., Gamba, P., Gualtieri, A., Tilton, J.: Recent advances in techniques for hyperspectral image processing. Remote. Sens. Environ. **113**, S110–S122 (2008)
16. Mountrakis, G., Ima, J., Ogole, C.: Support vector machines in remote sensing: a review. ISPRS J. Photogramm. Remote. Sens. **66**, 247–259 (2011)
17. Camps-Valls, G., Tuia, D., Bruzzone, L., Benediktsson, J.A.: Advances in hyperspectral image classification. IEEE Signal Process. Mag. **31**, 45–54 (2014)
18. Dorigo, W.A., Zurita-Milla, R., de Wit, A.J.W., Brazile, J., Singh, R., Schaepman, M.E.: A review on reflective remote sensing and data assimilation techniques for enhanced agroecosystem modeling. Int. J. Appl. Earth Obs. Geoinf. **9**, 165–193 (2007)
19. Schaepman, M., Ustin, S., Plaza, A., Painter, T., Verrelst, J., Liang, S.: Earth system science related imaging spectroscopy-an assessment. Remote. Sens. Environ. **113**, S123–S137 (2009)
20. Drusch, M., Del Bello, U., Carlier, S., Colin, O., Fernandez, V., Gascon, F., Hoersch, B., Isola, C., Laberinti, P., Martimort, P., Meygret, A., Spoto, F., Sy, O., Marchese, F., Bargellini, P.: Sentinel-2: ESA's optical high-resolution mission for GMES operational services. Remote. Sens. Environ. **120**, 25–36 (2012)
21. Donlon, C., Berruti, B., Buongiorno, A., Ferreira, M.H., Féménias, P., Frerick, J., Goryl, P., Klein, U., Laur, H., Mavrocordatos, C., Nieke, J., Rebhan, H., Seitz, B., Stroede, J., Sciarra,

R.: The global monitoring for environment and security (GMES) Sentinel-3 mission. Remote. Sens. Environ. **120**, 37–57 (2012)
22. Camps-Valls, G., Tuia, D., Gómez-Chova, L., Malo, J. (eds.): Remote Sensing Image Processing. Morgan & Claypool, San Rafael (2011)
23. Baret, F., Weiss, M., Lacaze, R., Camacho, F., Makhmara, H., Pacholcyzk, P., Smets, B.: Geov1: LAI and FAPAR essential climate variables and FCOVER global time series capitalizing over existing products. part1: principles of development and production. Remote. Sens. Environ. **137**, 299–309 (2013)
24. Beer, C., Reichstein, M., Tomelleri, E., Ciais, P., Jung, M., Carvalhais, N., Rödenbeck, C., Arain, M.A., Baldocchi, D., Bonan, G.B., Bondeau, A., Cescatti, A., Lasslop, G., Lindroth, A., Lomas, M., Luyssaert, S., Margolis, H., Oleson, K.W., Roupsard, O., Veenendaal, E., Viovy, N., Williams, C., Woodward, F.I., Papale, D.: Terrestrial gross carbon dioxide uptake: Global distribution and covariation with climate. Science **329**, 834 (2010)
25. Jung, M., Reichstein, M., Margolis, H.A., Cescatti, A., Richardson, A.D., Arain, M.A., Arneth, A., Bernhofer, C., Bonal, D., Chen, J., Gianelle, D., Gobron, N., Kiely, G., Kutsch, W., Lasslop, G., Law, B.E., Lindroth, A., Merbold, L., Montagnani, L., Moors, E.J., Papale, D., Sottocornola, M., Vaccari, F., Williams, C.: Global patterns of land-atmosphere fluxes of carbon dioxide, latent heat, and sensible heat derived from eddy covariance, satellite, and meteorological observations. J. Geophys. Res. Biogeosciences **116**, 1–16 (2011)
26. Sarker, L.R., Nichol, J.E.: Improved forest biomass estimates using ALOS AVNIR-2 texture indices. Remote. Sens. Environ. **115**, 968–977 (2011)
27. Guanter, L., Zhang, Y., Jung, M., Joiner, J., Voigt, M., Berry, J.A., Frankenberg, C., Huete, A., Zarco-Tejada, P., Lee, J.E., Moran, M.S., Ponce-Campos, G., Beer, C., Camps-Valls, G., Buchmann, N., Gianelle, D., Klumpp, K., Cescatti, A., Baker, J.M., Griffis, T.J.: Global and time-resolved monitoring of crop photosynthesis with chlorophyll fluorescence. Proc. Natl. Acad. Sci. PNAS **111**, E1327–E1333 (2014)
28. Camps-Valls, G., Gómez-Chova, L., Vila-Francés, J., Amorós-López, J., Muñoz-Marí, J., Calpe-Maravilla, J.: Retrieval of oceanic chlorophyll concentration with relevance vector machines. Remote. Sens. Environ. **105**, 23–33 (2006)
29. Yang, F., White, M., Michaelis, A., Ichii, K., Hashimoto, H., Votava, P., Zhu, A.X., Nemani, R.: Prediction of continental-scale evapotranspiration by combining MODIS and AmeriFlux data through support vector machine. IEEE Trans. Geosci. Remote. Sens. **44**, 3452–3461 (2006)
30. Durbha, S., King, R., Younan, N.: Support vector machines regression for retrieval of leaf area index from multiangle imaging spectroradiometer. Remote. Sens. Environ. **107**, 348–361 (2007)
31. Tuia, D., Verrelst, J., Alonso-Chordá, L., Pérez-Cruz, F., Camps-Valls, G.: Multioutput support vector regression for remote sensing biophysical parameter estimation. IEEE Geosci. Remote. Sens. Lett. **8**, 804–808 (2011)
32. Verrelst, J., Muñoz, J., Alonso, L., Delegido, J., Rivera, J., Moreno, J., Camps-Valls, G.: Machine learning regression algorithms for biophysical parameter retrieval: opportunities for Sentinel-2 and -3. Remote. Sens. Environ. **118**, 127–139 (2012)
33. Golub, G.H., Van Loan, C.F.: Matrix Computations. Johns Hopkins Studies in Mathematical Sciences. The Johns Hopkins University Press, Baltimore (1996)
34. Reed, M.C., Simon, B.: Functional Analysis. Methods of Modern Mathematical Physics, vol. I. Academic Press, New York (1980)
35. Schölkopf, B., Smola, A.: Learning with Kernels - Support Vector Machines, Regularization, Optimization and Beyond. MIT Press Series, Cambridge (2002)
36. Camps-Valls, G., Bruzzone, L. (eds.): Kernel Methods for Remote Sensing Data Analysis. Wiley, UK (2009)
37. Burges, C.J.C.: Geometry and invariance in kernel based methods. In: Schölkopf, B., Burges, C.J.C. (eds.) Advances in Kernel Methods - Support Vector Learning. MIT Press, Cambridge (1990)
38. Vapnik, V.N.: Statistical Learning Theory. Wiley, New York (1998)

39. Williams, C.K.I., Seeger, M.: Using the Nyström method to speed up kernel machines. In: Leen, T.K., Dietterich, T.G., Tresp, V. (eds.) Advances in Neural Information Processing Systems, NIPS2001. Vancouver, vol. 13, pp. 682–688. MIT Press, Canada (2001)
40. Hsieh, C.J., Si, S., Dhillon, I.S.: Fast prediction for large-scale kernel machines. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D., Weinberger, K.Q., (eds.) Advances in Neural Information Processing Systems, Curran Associates, Inc. pp. 3689–3697 (2014)
41. Rahimi, A., Recht, B.: Random features for large-scale kernel machines. In: Advances in Neural Information Processing Systems (2007)
42. Crammer, K., Singer, Y.: On the algorithmic implementation of multiclass kernel-based vector machines. J. Mach. Learn. Res. **2**, 265–292 (2002)
43. Gómez-Chova, L., Tuia, D., Moser, G., Camps-Valls, G.: Multimodal classification of remote sensing images: a review and future directions. Proc. IEEE **103**, 1560–1584 (2015)
44. Sonnenburg, S., Rätsch, G., Schafer, C., Schölkopf, B.: Large scale multiple kernel learning. J. Mach. Learn. Res. **7**, 1531–1565 (2006)
45. Tuia, D., Camps-Valls, G., Matasci, G., Kanevski, M.: Learning relevant image features with multiple kernel classification. IEEE Trans. Geosci. Remote. Sens. **48**, 3780–3791 (2010)
46. Gu, Y., Wang, S., Jia, X.: Spectral unmixing in multiple-kernel hilbert space for hyperspectral imagery. IEEE Trans. Geosci. Remote. Sens. **51**, 3968–3981 (2013)
47. Liu, K.H., Lin, Y.Y., Chen, C.S.: Linear spectral mixture analysis via multiple-kernel learning for hyperspectral image classification. IEEE Trans. Geosci. Remote. Sens. **53**, 2254–2269 (2015)
48. Gu, J., Jiao, L., Yang, S., Liu, F., Hou, B., Zhao, Z.: A multi-kernel joint sparse graph for SAR image segmentation. IEEE J. Sel. Top. Appl. Earth Obs. **9**, 1265–1285 (2016)
49. Gu, Y., Wang, C., You, D., Zhang, Y., Wang, S., Zhang, Y.: Representative multiple-kernel learning for classification of hyperspectral imagery. IEEE Trans. Geosci. Remote. Sens. **7**, 2852–2865 (2012)
50. Cusano, C., Napoletano, P., Schettini, R.: Remote sensing image classification exploiting multiple kernel learning. IEEE Geosci. Remote. Sens. Lett. **12**, 2331–2335 (2015)
51. Gu, Y., Gao, G., Zuo, D., You, D.: Model selection and classification with multiple kernel learning for hyperspectral images via sparsity. IEEE J. Sel. Top. Appl. Earth Obs. **7**, 2119–2130 (2014)
52. Wang, Q., Gu, Y., Tuia, D.: Discriminative multiple kernel learning for hyperspectral image classification. IEEE Trans. Geosci. Remote. Sens. **54**(7), 3912–3927 (2016)
53. Wang, L., Hao, S., Wang, Q., Atkinson, P.M.: A multiple-mapping kernel for hyperspectral image classification. IEEE Geosci. Remote. Sens. Lett. **12**, 978–982 (2015)
54. Zhang, Y., Yang, H.L., Prasad, S., Pasolli, E., Jung, J., Crawford, M.: Ensemble multiple kernel active learning for classification of multisource remote sensing data. IEEE J. Sel. Top. Appl. Earth Obs. **8**, 845–858 (2015)
55. Sun, Z., Wang, C., Wang, H., Li, J.: Learn multiple-kernel SVMs for domain adaptation in hyperspectral data. IEEE Geosci. Remote. Sens. Lett. **10**, 1224–1228 (2013)
56. Li, J., Marpu, P.R., Plaza, A., Bioucas-Dias, J., Benediktsson, J.A.: Generalized composite kernel framework for hyperspectral image classification. IEEE Trans. Geosci. Remote. Sens. **51**, 4816–4829 (2013)
57. Tuia, D., Camps-Valls, G.: Urban image classification with semisupervised multiscale cluster kernels. IEEE J. Sel. Top. Appl. Earth Obs. **4**, 65–74 (2011)
58. Rakotomamonjy, A., Bach, F., Canu, S., Grandvalet, Y.: SimpleMKL. J. Mach. Learn. Res. **9**, 2491–2521 (2008)
59. Barnsley, M., Settle, J., Cutter, M., Lobb, D., Teston, F.: The PROBA/CHRIS mission: a low-cost smallsat for hyperspectral, multi-angle, observations of the Earth surface and atmosphere. IEEE Trans. Geosci. Remote. Sens. **42**, 1512–1520 (2004)
60. Hajnsek, I., Bianchi, R., Davidson, M., Wooding, M.: The AgriSAR 2006 team: AgriSAR 2006 - Airborne SAR and optics campaigns for an improved monitoring of agricultural processes and practices. In: Fourth International Workshop on the Analysis of Multitemporal Remote Sensing Images. MultiTemp2007, Leuven, Belgium (2007)

61. Cristianini, N., Kandola, J., Elisseeff, A., Shawe-Taylor, J.: On kernel target alignment. Technical Report 2001-087, NeuroCOLT (2001)
62. Guanter, L., Richter, R., Kaufmann, H.: On the application of the MODTRAN4 atmospheric radiative transfer code to optical remote sensing. Int. J. Remote. Sens. **30**, 1407–1424 (2009)
63. Guanter, L., Ruiz-Verdú, A., Odermatt, D., Giardino, C., Simis, S., Estelles, V., Heege, T., Domínguez-Gómez, J.A., Moreno, J.: Atmospheric correction of ENVISAT/MERIS data over inland waters: validation for European lakes. Remote. Sens. Environ. **114**, 467–480 (2010)
64. Matasci, G., Longbotham, N., Pacifici, F,M,K., Tuia, D.: Understanding angular effects in VHR imagery and their significance for urban land-cover model portability: a study of two multi-angle in-track image sequences. ISPRS J. Int. Soc. Photogramm. Remote. Sens. **107**, 99–111 (2015)
65. Hong, G., Zhang, Y.: Radiometric normalization of IKONOS image using Quickbird image for urban area change detection. In: Proceedings of ISPRS 3rd International Symposium on Remote Sensing and Data Fusion Over Urban Areas, Tempe, AZ (2005)
66. Yang, Z., Mueller, R.: Heterogeneously sensed imagery radiometric response normalization for citrus grove change detection. In: Proceedings of SPIE Optics East, vol. 6761. Boston, MA (2007)
67. Tuia, D., Muñoz-Marí, J., Gómez-Chova, L., Malo, J.: Graph matching for adaptation in remote sensing. IEEE Trans. Geosci. Remote. Sens. **51**, 329–341 (2013)
68. Nielsen, A.A.: Multiset canonical correlations analysis and multispectral, truly multitemporal remote sensing data. IEEE Trans. Image Process. **11**, 293–305 (2002)
69. Volpi, M., Camps-Valls, G., Tuia, D.: Spectral alignment of cross-sensor images with automated kernel canonical correlation analysis. ISPRS J. Int. Soc. Photogramm. Remote. Sens. **107**, 50–63 (2015)
70. Wang, C., Krafft, P., Mahadevan, S.: Manifold alignment. In: Ma, Y., Fu, Y. (eds.) Manifold Learning: Theory and Applications. CRC Press, Boca Raton (2011)
71. Wang, C., Mahadevan, S.: Heterogeneous domain adaptation using manifold alignment. In: International Joint Conference on Artificial Intelligence (IJCAI) (2011)
72. Tuia, D., Volpi, M., Trolliet, M., Camps-Valls, G.: Semisupervised manifold alignment of multimodal remote sensing images. IEEE Trans. Geosci. Remote. Sens. **52**, 7708–7720 (2014)
73. Yang, H., Crawford, M.: Spectral and spatial proximity-based manifold alignment for multitemporal hyperspectral image classification. IEEE Trans. Geosci. Remote. Sens. **54**, 51–64 (2016)
74. Yang, H., Crawford, M.: Domain adaptation with preservation of manifold geometry for hyperspectral image classification. IEEE J. Sel. Top. Appl. Earth Obs. **9**, 543–555 (2016)
75. Tuia, D., Marcos, D., Camps-Valls, G.: Multi-temporal and multi-source remote sensing image classification by nonlinear relative normalization. ISPRS J. Int. Soc. Photo. Remote Sens. **120**, 1–12 (2016)
76. Liao, D., Qian, D., Zhou, J., Tang, Y.: A manifold alignment approach for hyperspectral image visualization with natural color. IEEE Trans. Geosci. Remote. Sens. **54**, 3151–3162 (2016)
77. Tuia, D., Camps-Valls, G.: Kernel manifold alignment for domain adaptation. PLoS One **11**, e0148655 (2016)
78. Schindler, K.: An overview and comparison of smooth labeling methods for land-cover classification. IEEE Trans. Geosci. Remote. Sens. **50**, 4534–4545 (2012)
79. Tuia, D., Volpi, M., Moser, G.: Getting pixels and regions to agree with conditional random fields. In: IEEE International Geoscience and Remote Sensing Symposium, IGARSS, Beijing, China (2016)
80. Moser, G., Serpico, S.B.: Combining support vector machines and markov random fields in an integrated framework for contextual image classification. IEEE Trans. Geosci. Remote. Sens. **51**, 2734–2752 (2013)
81. Volpi, M., Ferrari, V.: Structured prediction for urban scene semantic segmentation with geographic context. In: Joint Urban Remote Sensing Event (JURSE), Lausanne, Switzerland (2015)

82. Tuia, D., Muñoz-Marí, J., Kanevski, M., Camps-Valls, G.: Structured output SVM for remote sensing image classification. J. Signal Proc. Sys. **65**, 457–468 (2011)
83. Volpi, M., Ferrari, V.: Semantic segmentation of urban scenes by learning local class interactions. In: IEEE CVPR Workshop "Looking from above: when Earth observation meets vision", Boston, MA (2015)
84. Li, W., Du, Q., Xiong, M.: Kernel collaborative representation with tikhonov regularization for hyperspectral image classification. IEEE Geosci. Remote. Sens. Lett. **12**, 48–52 (2015)
85. Liu, J., Wu, Z., Li, J., Plaza, A., Yuan, Y.: Probabilistic-kernel collaborative representation for spatial-spectral hyperspectral image classification. IEEE Trans. Geosci. Remote. Sens. **54**, 2371–2384 (2016)
86. de Morsier, F., Borgeaud, M., Gass, V., Thiran, J.P., Tuia, D.: Kernel low-rank and sparse graph for unsupervised and semi-supervised classification of hyperspectral images. IEEE Trans. Geosci. Remote. Sens. **54**, 3410–3420 (2016)
87. Camps-Valls, G., Verrelst, J., Muoz-Mar, J., Laparra, V., Mateo-Jiménez, F., Gomez-Dans, J.: A survey on Gaussian processes for earth observation data analysis. IEEE Geosci. Remote. Sens. Mag. **4**, 58–78 (2016)
88. Hastie, T., Tibshirani, R., Friedman, J.H.: The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd edn. Springer, New York (2009)
89. Sampson, P., Guttorp, P.: Nonparametric estimation of nonstationary spatial covariance structure. J. Am. Stat. Assoc. Publ. **87**, 108–119 (1992)
90. Camps-Valls, G., Martínez-Ramón, M., Rojo-Álvarez, J.L., Muñoz-Marí, J.: Non-linear system identification with composite relevance vector machines. IEEE Signal Proc. Lett. **14**, 279–282 (2007)
91. Álvarez, M.A., Luengo, D., Lawrence, N.D.: Linear latent force models using gaussian processes. IEEE Trans. Pattern Anal. Mach. Intell. **35**, 2693–2705 (2013)
92. Smola, A.J., Schölkopf, B.: A tutorial on support vector regression. Stat. Comput. **14**, 199–222 (2004)
93. Tipping, M.: The relevance vector machine. In: Solla, S.A., Leen, T.K., Müller, K.R. (eds.) Advances in Neural Information Processing Systems 12. MIT Press, Cambridge (2000)
94. Verrelst, J., Alonso, L., Camps-Valls, G., Delegido, J., Moreno, J.: Retrieval of vegetation biophysical parameters using Gaussian process techniques. IEEE Trans. Geosci. Remote. Sens. **50**, 1832–1843 (2012)
95. Verrelst, J., Alonso, L., Rivera Caicedo, J., Moreno, J., Camps-Valls, G.: Gaussian process retrieval of chlorophyll content from imaging spectroscopy data. IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens. **6**, 867–874 (2013)
96. Bazi, Y., Melgani, F.: Toward an optimal svm classification system for hyperspectral remote sensing images. IEEE Trans. Geosci. Remote. Sens. **44**, 3374–3385 (2006)
97. Archibald, R., Fann, G.: Feature selection and classification of hyperspectral images with support vector machines. IEEE Geosci. Remote. Sens. Lett. **4**, 674–677 (2007)
98. Pal, M., Foody, G.: Feature selection for classification of hyperspectral data by SVM. IEEE Trans. Geosci. Remote. Sens. **48**, 2297–2307 (2010)
99. Verrelst, J., Rivera, J., Veroustraete, F., Muñoz Marí, J., Clevers, J., Camps-Valls, G., Moreno, J.: Experimental Sentinel-2 LAI estimation using parametric, non-parametric and physical retrieval methods - a comparison. ISPRS J. Int. Soc. Photogramm. Remote. Sens. **108**, 260–272 (2015)
100. Van Wittenberghe, S., Verrelst, J., Rivera, J., Alonso, L., Moreno, J., Samson, R.: Gaussian processes retrieval of leaf parameters from a multi-species reflectance, absorbance and fluorescence dataset. J. Photochem. Photobiol. B Biol. **134**, 37–48 (2014)
101. Verrelst, J., Rivera, J.G., Gitelson, A., Delegido, J., Moreno, J., Camps-Valls, G.: Spectral band selection for vegetation properties retrieval using Gaussian processes regression. Int. J. Appl. Earth Obs. Geoinf. **52**, 554–567 (2016)
102. Rivera Caicedo, J., Verrelst, J., Muñoz-Marí, J., Moreno, J., Camps-Valls, G.: Toward a semiautomatic machine learning retrieval of biophysical parameters. IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens. **7**, 1249–1259 (2014)

103. Jagermeyr, J., Gerten, D., Lucht, W., Hostert, P., Migliavacca, M., Nemani, R.: A high-resolution approach to estimating ecosystem respiration at continental scales using operational satellite data. Glob. Chang. Biol. **20**, 1191–1210 (2014)
104. Verrelst, J., Rivera, J., Moreno, J., Camps-Valls, G.: Gaussian processes uncertainty estimates in experimental Sentinel-2 LAI and leaf chlorophyll content retrieval. ISPRS J. Int. Soc. Photogramm. Remote. Sens. **86**, 157–167 (2013)
105. Campos-Taberner, M., García-Haro, F., Moreno, A., Gilabert, M., Sánchez-Ruiz, S., Martínez, B., Camps-Valls, G.: Mapping leaf area index with a smartphone and Gaussian processes. IEEE Geosci. Remote. Sens. Lett. **12**, 2501–2505 (2015)
106. O'Hagan, A.: Bayesian analysis of computer code outputs: a tutorial. Reliab. Eng. Syst. Saf. **91**, 1290–1300 (2006)
107. Kennedy, M., O'Hagan, A.: Bayesian calibration of computer models. J. R. Stat. Soc. Ser. B (Stat. Methodol.) **63**, 425–450 (2001)
108. Conti, S., Gosling, J., Oakley, J., O'Hagan, A.: Gaussian process emulation of dynamic computer codes. Biometrika **96**, 663–676 (2009)
109. Petropoulos, G., Wooster, M., Carlson, T., Kennedy, M., Scholze, M.: A global Bayesian sensitivity analysis of the 1D simsphere soil vegetation atmospheric transfer (SVAT) model using Gaussian model emulation. Ecol. Model. **220**, 2427–2440 (2009)
110. Castelletti, A., Galelli, S., Ratto, M., Soncini-Sessa, R., Young, P.: A general framework for dynamic emulation modelling in environmental problems. Environ. Model. Softw. **34**, 5–18 (2012)
111. Bounceur, N., Crucifix, M., Wilkinson, R., et al.: Global sensitivity analysis of the climate-vegetation system to astronomical forcing: an emulator-based approach. Earth Syst. Dyn. Discuss. **5**, 901–943 (2014)
112. Rivera, J.P., Verrelst, J., Gómez-Dans, J., Muñoz Marí, J., Moreno, J., Camps-Valls, G.: An emulator toolbox to approximate radiative transfer models with statistical learning. Remote. Sens. **7**, 9347 (2015)
113. Gómez-Dans, J.L., Lewis, P.E., Disney, M.: Efficient emulation of radiative transfer codes using Gaussian processes and application to land surface parameter inferences. Remote. Sens. **8**, 119 (2016)
114. Verrelst, J., Sabater, N., Rivera, J.P., Muñoz-Marí, J., Vicent, J., Camps-Valls, G., Moreno, J.: Emulation of leaf, canopy and atmosphere radiative transfer models for fast global sensitivity analysis. Remote. Sens. **8**, 673 (2016)