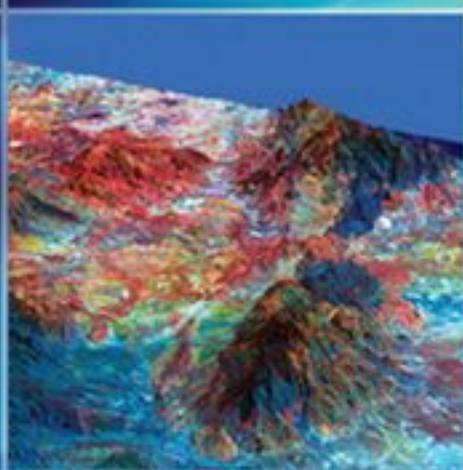
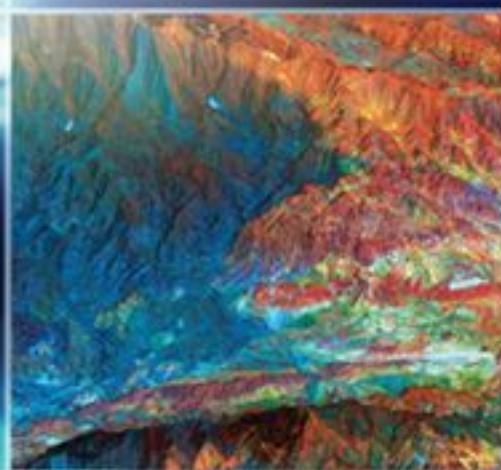


JIAN GUO LIU | PHILIPPA J MASON

# ESSENTIAL IMAGE PROCESSING AND GIS FOR REMOTE SENSING



# **Essential Image Processing and GIS for Remote Sensing**

**Jian Guo Liu  
Philippa J. Mason**

*Imperial College London, UK*



A John Wiley & Sons, Ltd., Publication



# **Essential Image Processing and GIS for Remote Sensing**



# **Essential Image Processing and GIS for Remote Sensing**

**Jian Guo Liu  
Philippa J. Mason**

*Imperial College London, UK*



A John Wiley & Sons, Ltd., Publication

This edition first published 2009, © 2009 by John Wiley & Sons Ltd.

Wiley-Blackwell is an imprint of John Wiley & Sons, formed by the merger of Wiley's global Scientific, Technical and Medical business with Blackwell Publishing.

*Registered office:* John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK

*Other Editorial Offices:*

9600 Garsington Road, Oxford, OX4 2DQ, UK  
111 River Street, Hoboken, NJ 07030-5774, USA

For details of our global editorial offices, for customer services and for information about how to apply for permission to reuse the copyright material in this book please see our website at [www.wiley.com/wiley-blackwell](http://www.wiley.com/wiley-blackwell)

The right of the author to be identified as the author of this work has been asserted in accordance with the Copyright, Designs and Patents Act 1988.

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, except as permitted by the UK Copyright, Designs and Patents Act 1988, without the prior permission of the publisher.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic books.

Designations used by companies to distinguish their products are often claimed as trademarks. All brand names and product names used in this book are trade names, service marks, trademarks or registered trademarks of their respective owners. The publisher is not associated with any product or vendor mentioned in this book. This publication is designed to provide accurate and authoritative information in regard to the subject matter covered. It is sold on the understanding that the publisher is not engaged in rendering professional services. If professional advice or other expert assistance is required, the services of a competent professional should be sought.

*Library of Congress Cataloguing-in-Publication Data*

Liu, Jian-Guo.

Essential image processing and GIS for remote sensing / Jian Guo Liu,  
Philippa J. Mason.

p. cm.

Includes index.

ISBN 978-0-470-51032-2 (HB) – ISBN 978-0-470-51031-5 (PB)

1. Remote sensing. 2. Geographic information systems. 3. Image processing.
4. Earth-Surface-Remote sensing. I. Mason, Philippa J. II. Title.

G70.4.L583 2009

621.36'78–dc22

2009007663

ISBN: 978-0-470-51032-2 (HB)  
978-0-470-51031-5 (PB)

A catalogue record for this book is available from the British Library.

Set in 10/12pt Times by Thomson Digital, Noida, India.

Printed in Singapore by Fabulous Printers Pte Ltd.

First Impression 2009

# Contents

---

<b>Overview of the Book</b>	<b>xv</b>
<b>Part One Image Processing</b>	<b>1</b>
<b>1 Digital Image and Display</b>	<b>3</b>
1.1 What is a digital image?	3
1.2 Digital image display	4
1.2.1 Monochromatic display	4
1.2.2 Tristimulus colour theory and RGB colour display	5
1.2.3 Pseudo colour display	7
1.3 Some key points	8
Questions	8
<b>2 Point Operations (Contrast Enhancement)</b>	<b>9</b>
2.1 Histogram modification and lookup table	9
2.2 Linear contrast enhancement	11
2.2.1 Derivation of a linear function from two points	12
2.3 Logarithmic and exponential contrast enhancement	13
2.3.1 Logarithmic contrast enhancement	13
2.3.2 Exponential contrast enhancement	14
2.4 Histogram equalization	14
2.5 Histogram matching and Gaussian stretch	15
2.6 Balance contrast enhancement technique	16
2.6.1 *Derivation of coefficients, $a$ , $b$ and $c$ for a BCET parabolic function	16
2.7 Clipping in contrast enhancement	18
2.8 Tips for interactive contrast enhancement	18
Questions	19
<b>3 Algebraic Operations (Multi-image Point Operations)</b>	<b>21</b>
3.1 Image addition	21
3.2 Image subtraction (differencing)	22
3.3 Image multiplication	22
3.4 Image division (ratio)	24

3.5	Index derivation and supervised enhancement	26
3.5.1	Vegetation indices	27
3.5.2	Iron oxide ratio index	28
3.5.3	TM clay (hydrated) mineral ratio index	29
3.6	Standardization and logarithmic residual	29
3.7	Simulated reflectance	29
3.7.1	Analysis of solar radiation balance and simulated irradiance	29
3.7.2	Simulated spectral reflectance image	30
3.7.3	Calculation of weights	31
3.7.4	Example: ATM simulated reflectance colour composite	32
3.7.5	Comparison with ratio and logarithmic residual techniques	33
3.8	Summary	34
	Questions	35
<b>4</b>	<b>Filtering and Neighbourhood Processing</b>	<b>37</b>
4.1	Fourier transform: understanding filtering in image frequency	37
4.2	Concepts of convolution for image filtering	39
4.3	Low-pass filters (smoothing)	40
4.3.1	Gaussian filter	41
4.3.2	The $k$ nearest mean filter	42
4.3.3	Median filter	42
4.3.4	Adaptive median filter	42
4.3.5	The $k$ nearest median filter	43
4.3.6	Mode (majority) filter	43
4.3.7	Conditional smoothing filter	43
4.4	High-pass filters (edge enhancement)	44
4.4.1	Gradient filters	45
4.4.2	Laplacian filters	46
4.4.3	Edge-sharpening filters	47
4.5	Local contrast enhancement	48
4.6	*FFT selective and adaptive filtering	48
4.6.1	FFT selective filtering	49
4.6.2	FFT adaptive filtering	51
4.7	Summary	54
	Questions	54
<b>5</b>	<b>RGB–IHS Transformation</b>	<b>57</b>
5.1	Colour coordinate transformation	57
5.2	IHS decorrelation stretch	59
5.3	Direct decorrelation stretch technique	61
5.4	Hue RGB colour composites	63
5.5	*Derivation of RGB–IHS and IHS–RGB transformations based on 3D geometry of the RGB colour cube	65
5.5.1	Derivation of RGB–IHS Transformation	65
5.5.2	Derivation of IHS–RGB transformation	66
5.6	*Mathematical proof of DDS and its properties	67
5.6.1	Mathematical proof of DDS	67
5.6.2	The properties of DDS	68
5.7	Summary	70
	Questions	70

<b>6 Image Fusion Techniques</b>	<b>71</b>
6.1 RGB–IHS transformation as a tool for data fusion	71
6.2 Brovey transform (intensity modulation)	73
6.3 Smoothing-filter-based intensity modulation	73
6.3.1 The principle of SFIM	74
6.3.2 Merits and limitation of SFIM	75
6.4 Summary	76
Questions	76
<b>7 Principal Component Analysis</b>	<b>77</b>
7.1 Principle of PCA	77
7.2 Principal component images and colour composition	80
7.3 Selective PCA for PC colour composition	82
7.3.1 Dimensionality and colour confusion reduction	82
7.3.2 Spectral contrast mapping	83
7.3.3 FPCS spectral contrast mapping	84
7.4 Decorrelation stretch	85
7.5 Physical-property-orientated coordinate transformation and tasseled cap transformation	85
7.6 Statistic methods for band selection	88
7.6.1 Review of Chavez <i>et al.</i> 's and Sheffield's methods	88
7.6.2 Index of three-dimensionality	89
7.7 Remarks	89
Questions	90
<b>8 Image Classification</b>	<b>91</b>
8.1 Approaches of statistical classification	91
8.1.1 Unsupervised classification	91
8.1.2 Supervised classification	91
8.1.3 Classification processing and implementation	92
8.1.4 Summary of classification approaches	92
8.2 Unsupervised classification (iterative clustering)	92
8.2.1 Iterative clustering algorithms	92
8.2.2 Feature space iterative clustering	93
8.2.3 Seed selection	94
8.2.4 Cluster splitting along PC1	95
8.3 Supervised classification	96
8.3.1 Generic algorithm of supervised classification	96
8.3.2 Spectral angle mapping classification	96
8.4 Decision rules: dissimilarity functions	97
8.4.1 Box classifier	97
8.4.2 Euclidean distance: simplified maximum likelihood	98
8.4.3 Maximum likelihood	98
8.4.4 *Optimal multiple point reassignment	98
8.5 Post-classification processing: smoothing and accuracy assessment	99
8.5.1 Class smoothing process	99
8.5.2 Classification accuracy assessment	100
8.6 Summary	102
Questions	102

<b>9 Image Geometric Operations</b>	<b>105</b>
9.1 Image geometric deformation	105
9.1.1 Platform flight coordinates, sensor status and imaging geometry	105
9.1.2 Earth rotation and curvature	107
9.2 Polynomial deformation model and image warping co-registration	108
9.2.1 Derivation of deformation model	109
9.2.2 Pixel DN resampling	110
9.3 GCP selection and automation	111
9.3.1 Manual and semi-automatic GCP selection	111
9.3.2 *Towards automatic GCP selection	111
9.4 *Optical flow image co-registration to sub-pixel accuracy	113
9.4.1 Basics of phase correlation	113
9.4.2 Basic scheme of pixel-to-pixel image co-registration	114
9.4.3 The median shift propagation technique	115
9.4.4 Summary of the refined pixel-to-pixel image co-registration and assessment	117
9.5 Summary	118
Questions	119
<b>10 *Introduction to Interferometric Synthetic Aperture Radar Techniques</b>	<b>121</b>
10.1 The principle of a radar interferometer	121
10.2 Radar interferogram and DEM	123
10.3 Differential InSAR and deformation measurement	125
10.4 Multi-temporal coherence image and random change detection	127
10.5 Spatial decorrelation and ratio coherence technique	129
10.6 Fringe smoothing filter	132
10.7 Summary	132
Questions	134
<b>Part Two Geographical Information Systems</b>	<b>135</b>
<b>11 Geographical Information Systems</b>	<b>137</b>
11.1 Introduction	137
11.2 Software tools	138
11.3 GIS, cartography and thematic mapping	138
11.4 Standards, interoperability and metadata	139
11.5 GIS and the Internet	140
<b>12 Data Models and Structures</b>	<b>141</b>
12.1 Introducing spatial data in representing geographic features	141
12.2 How are spatial data different from other digital data?	141
12.3 Attributes and measurement scales	142
12.4 Fundamental data structures	143
12.5 Raster data	143
12.5.1 Data quantization and storage	143
12.5.2 Spatial variability	145
12.5.3 Representing spatial relationships	145

12.5.4 The effect of resolution	146
12.5.5 Representing surfaces	147
12.6 Vector data	147
12.6.1 Representing logical relationships	148
12.6.2 Extending the vector data model	153
12.6.3 Representing surfaces	155
12.7 Conversion between data models and structures	157
12.7.1 Vector to raster conversion (rasterization)	158
12.7.2 Raster to vector conversion (vectorization)	160
12.8 Summary	161
Questions	162
<b>13 Defining a Coordinate Space</b>	<b>163</b>
13.1 Introduction	163
13.2 Datums and projections	163
13.2.1 Describing and measuring the Earth	164
13.2.2 Measuring height: the geoid	165
13.2.3 Coordinate systems	166
13.2.4 Datums	166
13.2.5 Geometric distortions and projection models	167
13.2.6 Major map projections	169
13.2.7 Projection specification	172
13.3 How coordinate information is stored and accessed	173
13.4 Selecting appropriate coordinate systems	174
Questions	175
<b>14 Operations</b>	<b>177</b>
14.1 Introducing operations on spatial data	177
14.2 Map algebra concepts	178
14.2.1 Working with null data	178
14.2.2 Logical and conditional processing	179
14.2.3 Other types of operator	179
14.3 Local operations	181
14.3.1 Primary operations	181
14.3.2 Unary operations	182
14.3.3 Binary operations	184
14.3.4 N-ary operations	185
14.4 Neighbourhood operations	185
14.4.1 Local neighbourhood	185
14.4.2 Extended neighbourhood	191
14.5 Vector equivalents to raster map algebra	192
14.6 Summary	194
Questions	195
<b>15 Extracting Information from Point Data: Geostatistics</b>	<b>197</b>
15.1 Introduction	197
15.2 Understanding the data	198
15.2.1 Histograms	198

15.2.2 Spatial autocorrelation	198
15.2.3 Variograms	199
15.2.4 Underlying trends and natural barriers	200
15.3 Interpolation	201
15.3.1 Selecting sample size	201
15.3.2 Interpolation methods	202
15.3.3 Deterministic interpolators	202
15.3.4 Stochastic interpolators	207
15.4 Summary	209
Questions	209
<b>16 Representing and Exploiting Surfaces</b>	<b>211</b>
16.1 Introduction	211
16.2 Sources and uses of surface data	211
16.2.1 Digital elevation models	211
16.2.2 Vector surfaces and objects	214
16.2.3 Uses of surface data	215
16.3 Visualizing surfaces	215
16.3.1 Visualizing in two dimensions	216
16.3.2 Visualizing in three dimensions	218
16.4 Extracting surface parameters	220
16.4.1 Slope: gradient and aspect	220
16.4.2 Curvature	222
16.4.3 Surface topology: drainage networks and watersheds	225
16.4.4 Viewshed	226
16.4.5 Calculating volume	228
16.5 Summary	229
Questions	229
<b>17 Decision Support and Uncertainty</b>	<b>231</b>
17.1 Introduction	231
17.2 Decision support	231
17.3 Uncertainty	232
17.3.1 Criterion uncertainty	233
17.3.2 Threshold uncertainty	233
17.3.3 Decision rule uncertainty	234
17.4 Risk and hazard	234
17.5 Dealing with uncertainty in spatial analysis	235
17.5.1 Error assessment (criterion uncertainty)	235
17.5.2 Fuzzy membership (threshold uncertainty)	236
17.5.3 Multi-criteria decision making (decision rule uncertainty)	236
17.5.4 Error propagation and sensitivity analysis (decision rule uncertainty)	237
17.5.5 Result validation (decision rule uncertainty)	238
17.6 Summary	239
Questions	239
<b>18 Complex Problems and Multi-Criteria Evaluation</b>	<b>241</b>
18.1 Introduction	241
18.2 Different approaches and models	242

18.2.1 Knowledge-driven approach (conceptual)	242
18.2.2 Data-driven approach (empirical)	242
18.2.3 Data-driven approach (neural network)	243
18.3 Evaluation criteria	243
18.4 Deriving weighting coefficients	244
18.4.1 Rating	244
18.4.2 Ranking	245
18.4.3 Pairwise comparison	245
18.5 Multi-criteria combination methods	248
18.5.1 Boolean logical combination	248
18.5.2 Index-overlay and algebraic combination	248
18.5.3 Weights of evidence modelling based on bayesian probability theory	249
18.5.4 Belief and Dempster–Shafer theory	251
18.5.5 Weighted factors in linear combination	252
18.5.6 Fuzzy logic	254
18.5.7 Vectorial fuzzy modelling	256
18.6 Summary	258
Questions	258
<b>Part Three Remote Sensing Applications</b>	<b>259</b>
<b>19 Image Processing and GIS Operation Strategy</b>	<b>261</b>
19.1 General image processing strategy	262
19.1.1 Preparation of basic working dataset	263
19.1.2 Image processing	266
19.1.3 Image interpretation and map composition	270
19.2 Remote-sensing-based GIS projects: from images to thematic mapping	271
19.3 An example of thematic mapping based on optimal visualization and interpretation of multi-spectral satellite imagery	272
19.3.1 Background information	272
19.3.2 Image enhancement for visual observation	274
19.3.3 Data capture and image interpretation	274
19.3.4 Map composition	278
19.4 Summary	279
Questions	280
<b>20 Thematic Teaching Case Studies in SE Spain</b>	<b>281</b>
20.1 Thematic information extraction (1): gypsum natural outcrop mapping and quarry change assessment	281
20.1.1 Data preparation and general visualization	281
20.1.2 Gypsum enhancement and extraction based on spectral analysis	283
20.1.3 Gypsum quarry changes during 1984–2000	284
20.1.4 Summary of the case study	287
20.2 Thematic information extraction (2): spectral enhancement and mineral mapping of epithermal gold alteration, and iron ore deposits in ferroan dolomite	287
20.2.1 Image datasets and data preparation	287
20.2.2 ASTER image processing and analysis for regional prospectivity	288

20.2.3 ATM image processing and analysis for target extraction	292
20.2.4 Summary	296
20.3 Remote sensing and GIS: evaluating vegetation and land-use change in the Nijar Basin, SE Spain	296
20.3.1 Introduction	296
20.3.2 Data preparation	297
20.3.3 Highlighting vegetation	298
20.3.4 Highlighting plastic greenhouses	300
20.3.5 Identifying change between different dates of observation	302
20.3.6 Summary	304
20.4 Applied remote sensing and GIS: a combined interpretive tool for regional tectonics, drainage and water resources	304
20.4.1 Introduction	304
20.4.2 Geological and hydrological setting	305
20.4.3 Case study objectives	306
20.4.4 Land use and vegetation	307
20.4.5 Lithological enhancement and discrimination	310
20.4.6 Structural enhancement and interpretation	313
20.4.7 Summary	318
Questions	320
References	321
<b>21 Research Case Studies</b>	<b>323</b>
21.1 Vegetation change in the three parallel rivers region, Yunnan province, China	323
21.1.1 Introduction	323
21.1.2 The study area and data	324
21.1.3 Methodology	324
21.1.4 Data processing	326
21.1.5 Interpretation of regional vegetation changes	328
21.1.6 Summary	332
21.2 Landslide hazard assessment in the three gorges area of the Yangtze river using ASTER imagery: Wushan–Badong–Zogui	334
21.2.1 Introduction	334
21.2.2 The study area	334
21.2.3 Methodology: multi-variable elimination and characterization	336
21.2.4 Terrestrial information extraction	339
21.2.5 DEM and topographic information extraction	344
21.2.6 Landslide hazard mapping	347
21.2.7 Summary	349
21.3 Predicting landslides using fuzzy geohazard mapping; an example from Piemonte, North-west Italy	350
21.3.1 Introduction	350
21.3.2 The study area	352
21.3.3 A holistic GIS-based approach to landslide hazard assessment	354
21.3.4 Summary	357
21.4 Land surface change detection in a desert area in Algeria using multi-temporal ERS SAR coherence images	359
21.4.1 The study area	359
21.4.2 Coherence image processing and evaluation	360

21.4.3 Image visualization and interpretation for change detection	361
21.4.4 Summary	366
Questions	366
References	366
<b>22 Industrial Case Studies</b>	<b>371</b>
22.1 Multi-criteria assessment of mineral prospectivity, in SE Greenland	371
22.1.1 Introduction and objectives	371
22.1.2 Area description	372
22.1.3 Litho-tectonic context – why the project’s concept works	373
22.1.4 Mineral deposit types evaluated	374
22.1.5 Data preparation	374
22.1.6 Multi-criteria spatial modelling	381
22.1.7 Summary	384
Acknowledgements	386
22.2 Water resource exploration in Somalia	386
22.2.1 Introduction	386
22.2.2 Data preparation	387
22.2.3 Preliminary geological enhancements and target area identification	388
22.2.4 Discrimination potential aquifer lithologies using ASTER spectral indices	390
22.2.5 Summary	397
Questions	397
References	397
<b>Part Four Summary</b>	<b>399</b>
<b>23 Concluding Remarks</b>	<b>401</b>
23.1 Image processing	401
23.2 Geographical information systems	404
23.3 Final remarks	407
<b>Appendix A: Imaging Sensor Systems and Remote Sensing Satellites</b>	<b>409</b>
A.1 Multi-spectral sensing	409
A.2 Broadband multi-spectral sensors	413
A.2.1 Digital camera	413
A.2.2 Across-track mechanical scanner	414
A.2.3 Along-track push-broom scanner	415
A.3 Thermal sensing and thermal infrared sensors	416
A.4 Hyperspectral sensors (imaging spectrometers)	417
A.5 Passive microwave sensors	418
A.6 Active sensing: SAR imaging systems	419
<b>Appendix B: Online Resources for Information, Software and Data</b>	<b>425</b>
B.1 Software – proprietary, low cost and free (shareware)	425
B.2 Information and technical information on standards, best practice, formats, techniques and various publications	426

B.3 Data sources including online satellite imagery from major suppliers, DEM data plus GIS maps and data of all kinds	426
<b>References</b>	<b>429</b>
General references	429
Image processing	429
GIS	430
Remote sensing	430
Part One References and further reading	430
Part Two References and further reading	433
<b>Index</b>	<b>437</b>

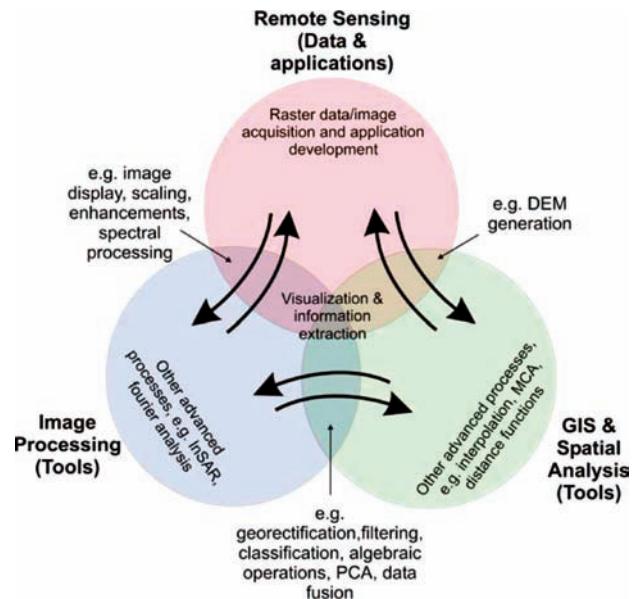
# Overview of the Book

From an applied viewpoint, and mainly for Earth observation, remote sensing is a tool for collecting raster data or images. Remotely sensed images represent an objective record of the spectrum relating to the physical properties and chemical composition of the Earth surface materials. Extracting information from images is, on the other hand, a subjective process. People with differing application foci will derive very different thematic information from the same source image. Image processing thus becomes a vital tool for the extraction of thematic and/or quantitative information from raw image data. For more comprehensive analysis, the images need to be analysed in conjunction with other complementary data, such as existing thematic maps of topography, geomorphology, geology and land use, or with geochemical and geophysical survey data, or ‘ground truth’ data, logistical and infrastructure information, which is where the geographical information system (GIS) comes into play. GIS contains highly sophisticated tools for the management, display and analysis of all kinds of spatially referenced information.

Remote sensing, image processing and GIS are all extremely broad subjects in their own right and are far too broad to be covered in one book. As illustrated in Figure 1, this book aims to pinpoint the overlap between the three subjects, providing an overview of essential techniques and a selection of case studies in a variety of application areas. The application cases are biased towards the earth

sciences but the image processing and GIS techniques are generic and therefore transferable skills suited to all applications.

In this book, we have presented a unique combination of tools, techniques and applications which we hope will be of use to a wide community of ‘geoscientists’ and ‘remote sensors’. The book begins in Part One with the fundamentals of the core image processing tools used in remote sensing and GIS with adequate mathematical details. It then becomes slightly more applied and less mathematical in Part Two to cover the wide scope of GIS where many of those core image processing tools are used in different contexts. Part Three contains the entirely applied part of the book where we describe a selection of cases where image processing and GIS have been used, by us, in teaching, research and industrial projects in which there is a dominant remote sensing component. The book has been written with university students and lecturers in mind as a principal textbook. For students’ needs in particular, we have tried to convey knowledge in simple words, with clear explanations and with conceptual illustrations. For image processing and GIS, mathematics is unavoidable, but we understand that this may be offputting for some. To minimize such effects, we try to emphasize the concepts, explaining in common-sense terms rather than in too much mathematical detail. The result is intended to be a comprehensive yet ‘easy learning’ solution to a fairly challenging topic.



**Figure 1** Schematic illustration of the scope of this book

On the other hand, the book indeed presents in depth some novel image processing techniques and GIS approaches. There are sections providing extended coverage of necessary mathematics and advanced materials for use by course tutors and lecturers; these sections will be marked by an

asterisk. Hence the book is for both students and teachers. With many of our developed techniques and most recent research case studies, it is also an excellent reference book for higher level readers including researchers and professionals in remote sensing application sectors.

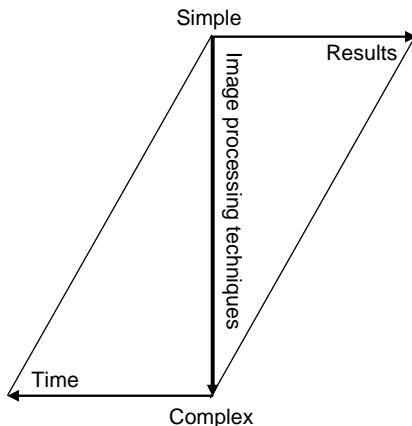
# Part One

## Image Processing

This part covers the most essential image processing techniques for image visualization, quantitative analysis and thematic information extraction for remote sensing applications. A series of chapters introduce topics with increasing complexity from basic visualization algorithms, which can be easily used to improve digital camera pictures, to much more complicated multi-dimensional transform-based techniques.

Digital image processing can improve image visual quality, selectively enhance and highlight particular image features and classify, identify and extract spectral and spatial patterns representing different phenomena from images. It can also arbitrarily change image geometry and illumination conditions to give different views of the same image. Importantly, *image processing cannot increase any information from the original image data*, although it can indeed optimize the visualization for us to see more information from the enhanced images than from the original.

For real applications our considered opinion, based on years of experience, is that *simplicity is beautiful*. Image processing does not follow the well-established physical law of energy conservation. As shown in Figure P.1, often the results produced using very simple processing techniques in the first 10 minutes of your project may actually represent 90% of the job done! This should not encourage you to abandon this book after the first three chapters, since it is the remaining 10% that you achieve during the 90% of your time that will serve the highest level objectives of your project. The key point is that thematic image processing should be application driven whereas our learning is usually technique driven.



**Figure P.1** This simple diagram is to illustrate that the image processing result is not necessarily proportional to the time/effort spent. On the one hand, you may spend little time in achieving the most useful results and with simple techniques; on the other hand, you may spend a lot of time achieving very little using complicated techniques



# 1

## Digital Image and Display

### 1.1 What is a digital image?

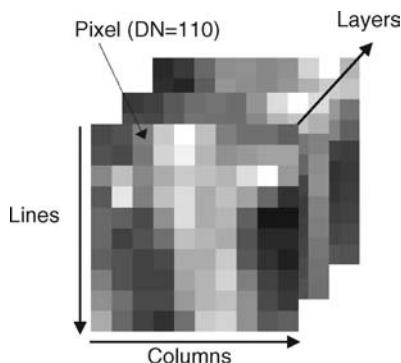
An image is a picture, photograph or any form of a two-dimensional representation of objects or a scene. The information in an image is presented in tones or colours. A *digital image* is a two-dimensional array of numbers. Each cell of a digital image is called a pixel and the number representing the brightness of the pixel is called a digital number (DN) (Figure 1.1). As a two-dimensional (2D) array, a digital image is composed of data in lines and columns. The position of a pixel is allocated with the line and column of its DN. Such regularly arranged data, without *x* and *y* coordinates, are usually called *raster data*. As digital images are nothing more than data arrays, mathematical operations can be readily performed on the digital numbers of images. Mathematical operations on digital images are called *digital image processing*.

Digital image data can also have a third dimension: *layers* (Figure 1.1). Layers are the images of the same scene but containing different information. In multi-spectral images, layers are the images of different spectral ranges called *bands* or *channels*. For instance, a colour picture taken by a digital camera is composed of three bands containing red, green and blue spectral information individually. The term ‘band’ is more often used than ‘layer’ to refer to multi-spectral images. Generally speaking, geometrically registered multi-dimensional datasets of the same scene can be considered as layers

of an image. For example, we can digitize a geological map and then co-register the digital map with a Landsat thematic mapper (TM) image. Then the digital map becomes an extra layer of the scene beside the seven TM spectral bands. Similarly, if we have a dataset of a digital elevation model (DEM) to which a SPOT image is rectified, then the DEM can be considered as a layer of the SPOT image beside its four spectral bands. In this sense, we can consider a set of co-registered digital images as a three-dimensional (3D) dataset and with the ‘third’ dimension providing the link between image processing and GIS.

A digital image can be stored as a file in a computer data store on a variety of media, such as a hard disk, CD, DVD or tape. It can be displayed in black and white or in colour on a computer monitor as well as in hard copy output such as film or print. It may also be output as a simple array of numbers for numerical analysis. As a digital image, its advantages include:

- The images do not change with environmental factors as hard copy pictures and photographs do.
- The images can be identically duplicated without any change or loss of information.
- The images can be mathematically processed to generate new images without altering the original images.
- The images can be electronically transmitted from or to remote locations without loss of information.



**Figure 1.1** A digital image and its elements

Remotely sensed images are acquired by sensor systems onboard aircraft or spacecraft, such as Earth observation satellites. The sensor systems can be categorized into two major branches: *passive sensors* and *active sensors*. Multi-spectral optical systems are passive sensors that use solar radiation as the principal source of illumination for imaging. Typical examples include across-track and push-broom multi-spectral scanners, and digital cameras. An active sensor system provides its own mean of illumination for imaging, such as synthetic aperture radar (SAR). Details of major remote sensing satellites and their sensor systems are beyond the scope of this book but we provide a summary in Appendix A for your reference.

## 1.2 Digital image display

We live in a world of colour. The colours of objects are the result of selective absorption and reflection of electromagnetic radiation from illumination sources. Perception by the human eye is limited to the spectral range of  $0.38\text{--}0.75\,\mu\text{m}$ , that is a very small part of the solar spectral range. The world is actually far more colourful than we can see. Remote sensing technology can record over a much wider spectral range than human visual ability and the resultant digital images can be displayed as either black and white or colour images using an electronic device such as a computer monitor. In digital image display, the tones or colours are visual representations of the image information recorded as digital image DNs, but they do not necessarily

convey the physical meanings of these DNs. We will explain this further in our discussion on false colour composites later.

The wavelengths of major spectral regions used for remote sensing are listed below:

Visible light (VIS)	0.4–0.7 $\mu\text{m}$
Blue (B)	0.4–0.5 $\mu\text{m}$
Green (G)	0.5–0.6 $\mu\text{m}$
Red (R)	0.6–0.7 $\mu\text{m}$
Visible–photographic infrared	0.5–0.9 $\mu\text{m}$
Reflective infrared (IR)	0.7–3.0 $\mu\text{m}$
Nearer infrared (NIR)	0.7–1.3 $\mu\text{m}$
Short-wave infrared (SWIR)	1.3–3.0 $\mu\text{m}$
Thermal infrared (TIR):	3–5 $\mu\text{m}$ , 8–14 $\mu\text{m}$
Microwave	0.1–100 cm

Commonly used abbreviations of the spectral ranges are denoted by the letters in brackets in the list above. The spectral range covering visible light and nearer infrared is the most popular for broadband multi-spectral sensor systems and it is usually denoted as VNIR.

### 1.2.1 Monochromatic display

Any image, either a panchromatic image or a spectral band of a multi-spectral image, can be displayed as a black and white (B/W) image by a monochromatic display. The display is implemented by converting DNs to electronic signals in a series of energy levels that generate different grey tones (brightness) from black to white, and thus formulate a B/W image display. Most image processing systems support an 8 bit graphical display, which corresponds to 256 grey levels, and displays DNs from 0 (black) to 255 (white). This display range is wide enough for human visual capability. It is also sufficient for some of the more commonly used remotely sensed images, such as Landsat TM/ETM+, SPOT HRV and Terra-1 ASTER VIR-SWIR (see Appendix A); the DN ranges of

these images are not wider than 0–255. On the other hand, many remotely sensed images have much wider DN ranges than 8 bits, such as those from Ikonos and Quickbird, whose images have an 11 bit DN range (0–2047). In this case, the images can still be visualized in an 8 bit display device in various ways, such as by compressing the DN range into 8 bits or displaying the image in scenes of several 8 bit intervals of the whole DN range. Many sensor systems offer wide dynamic ranges to ensure that the sensors can record across all levels of radiation energy without localized sensor adjustment. Since the received solar radiation does not normally vary significantly within an image scene of limited size, the actual DN range of the scene is usually much narrower than the full dynamic range of the sensor and thus can be well adapted into an 8 bit DN range for display.

In a monochromatic display of a spectral band image, the brightness (grey level) of a pixel is proportional to the reflected energy in this band from the corresponding ground area. For instance, in a B/W display of a red band image, light red appears brighter than dark red. This is also true for invisible bands (e.g. infrared bands), though the ‘colours’ cannot be seen. After all, any digital image is composed of DNs; the physical meaning of DNs depends on the source of the image. A monochromatic display visualizes DNs in grey tones from black to white, while ignoring the physical relevance.

### 1.2.2 Tristimulus colour theory and RGB colour display

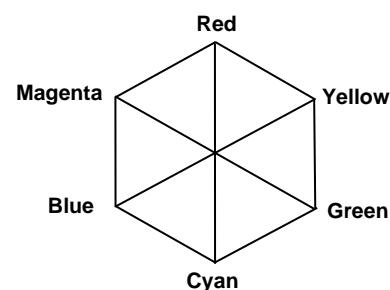
If you understand the structure and principle of a colour TV tube, you must know that the tube is composed of three colour guns of red, green and blue. These three colours are known as *primary colours*. The mixture of the light from these three primary colours can produce any colour on a TV. This property of the human perception of colour can be explained by the *tristimulus colour theory*. The human retina has three types of cones and the response by each type of cone is a function of the wavelength of the incident light; it peaks at 440 nm (blue), 545 nm (green) and 680 nm (red). In other

words, each type of cone is primarily sensitive to one of the primary colours: blue, green or red. A colour perceived by a person depends on the proportion of each of these three types of cones being stimulated and thus can be expressed as a triplet of numbers ( $r, g, b$ ) even though visible light is electromagnetic radiation in a continuous spectrum of 380–750 nm. A light of non-primary colour  $C$  will stimulate different portions of each cone type to form the perception of this colour:

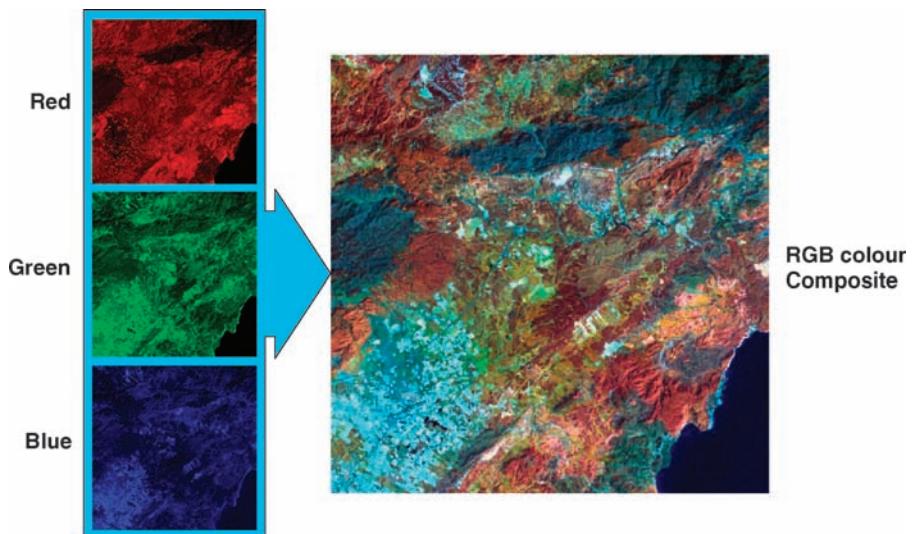
$$C = rR + gG + bB. \quad (1.1)$$

Equal mixtures of the three primary colours ( $r=g=b$ ) give white or grey, while equal mixtures of any two primary colours generate a complementary colour. As shown in Figure 1.2, the complementary colours of red, green and blue are cyan, magenta and yellow. The three complementary colours can also be used as primaries to generate various colours, as in colour printing. If you have experience of colour painting, you must know that any colour can be generated by mixing three colours: red, yellow and blue; this is based on the same principle.

Digital image colour display is based entirely on the tristimulus colour theory. A colour monitor, like a colour TV, is composed of three precisely registered colour guns, namely red, green and blue. In the red gun, pixels of an image are displayed in reds of different intensity (i.e. dark red, light red, etc.) depending on their DNs. The same is true of the green and blue guns. Thus if the red, green and blue bands of a multi-spectral image are displayed in red, green and blue simultaneously, a colour image is generated (Figure 1.3) in which the colour of a pixel is decided by the DNs of red, green and blue



**Figure 1.2** The relation of the primary colours to their complementary colours



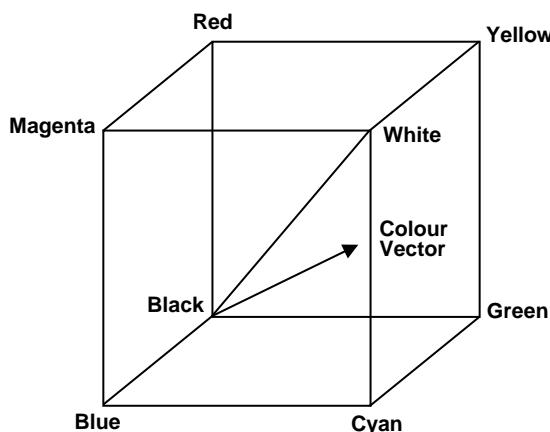
**Figure 1.3** Illustration of RGB additive colour image display

bands ( $r, g, b$ ). For instance, if a pixel has red and green DNs of 255 and blue DN of 0, it will appear in pure yellow on display. This kind colour display system is called an *additive RGB colour composite system*. In this system, different colours are generated by additive combinations of **Red**, **Green** and **Blue** components.

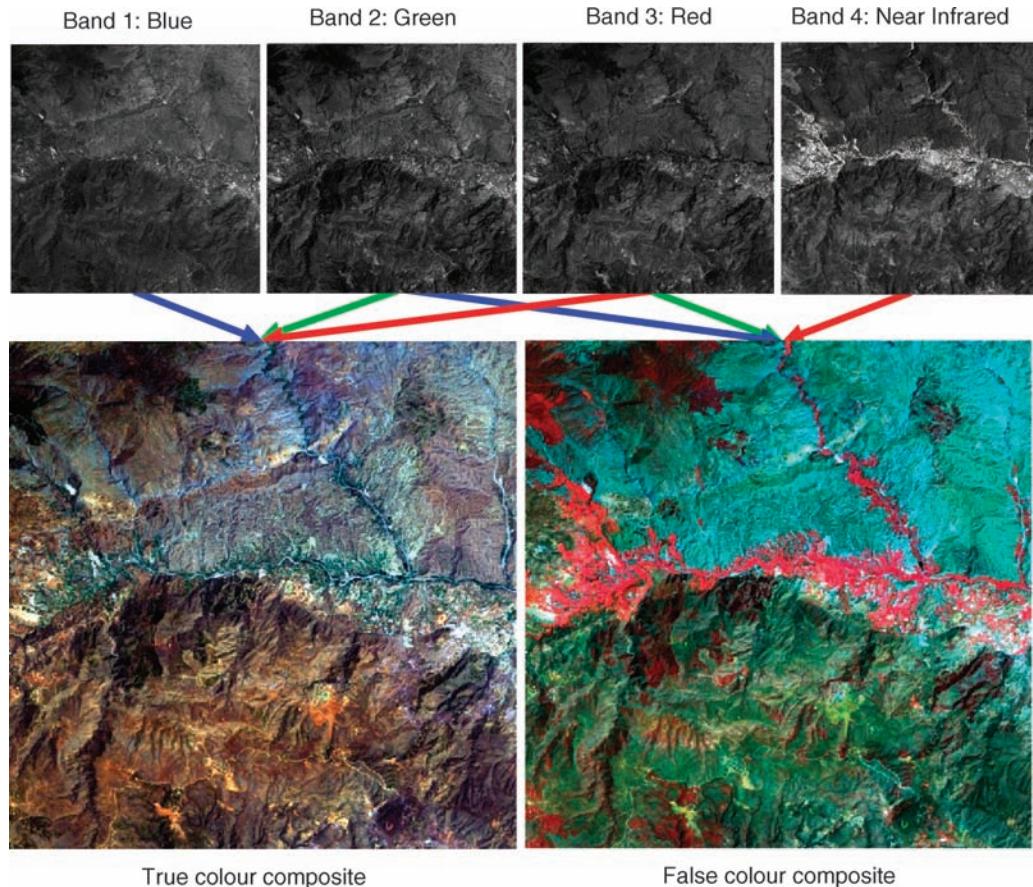
As shown in Figure 1.4, consider the components of an RGB display as the orthogonal axes of a 3D colour space; the maximum possible DN level in each component of the display defines the *RGB colour cube*. Any image pixel in this system may be

represented by a vector from the origin to somewhere within the colour cube. Most standard RGB display system can display 8 bits per pixel per channel, up to 24 bits =  $256^3$  different colours. This capacity is enough to generate a so-called ‘true colour’ image. The line from the origin of the colour cube to the opposite convex corner is known as the *grey line* because pixel vectors that lie on this line have equal components in red, green and blue (i.e.  $r = g = b$ ). If the same band is used as red, green and blue components, all the pixels will lie on the grey line. In this case, a B/W image will be produced even though a colour display system is used.

As mentioned before, although colours lie in the visible spectral range of 380–750 nm, they are used as a tool for information visualization in the colour display of all digital images. Thus, for digital image display, the assignment of each primary colour for a spectral band or layer can arbitrarily depend on the requirements of the application, which may not necessarily correspond to the actual colour of the spectral range of the band. If we display three image bands in the red, green and blue spectral ranges in RGB, then a *true colour composite* (TCC) image is generated (Figure 1.5, bottom left). Otherwise, if the image bands displayed in red, green and blue do not match the spectra of these three primary colours, a *false colour composite* (FCC) image is produced. A typical example is the so-called



**Figure 1.4** The RGB colour cube



**Figure 1.5** True colour and false colour composites of blue, green, red and near-infrared bands of a Landsat-7 ETM+ image. If we display the blue band in blue, green band in green and red band in red, then a true colour composite is produced as shown at the bottom left. If we display the green band in blue, red band in green and near-infrared band in red, then a so-called standard false colour composite is produced as shown at the bottom right

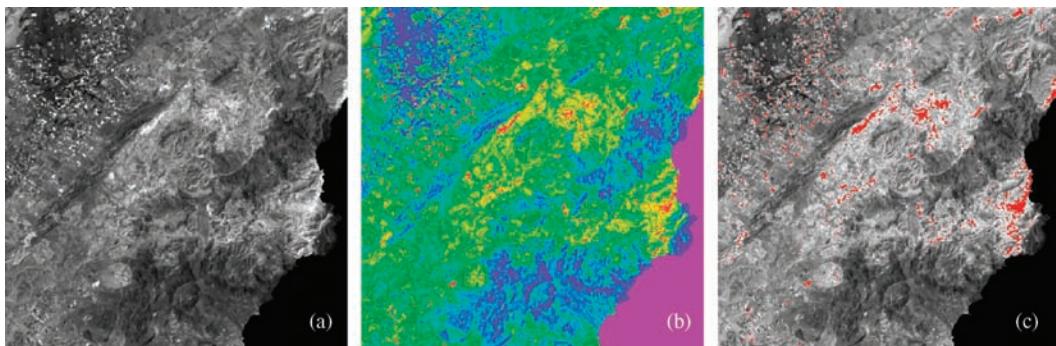
*standard false colour composite* (SFCC) in which the near-infrared band is displayed in red, the red band in green and the green band in blue (Figure 1.5, bottom right). The SFCC effectively highlights any vegetation distinctively in red. Obviously, we could display various image layers, which are without any spectral relevance, as a false colour composite. The false colour composite is the general case of an RGB colour display while the true colour composite is only a special case of it.

### 1.2.3 Pseudo colour display

The human eye can recognize far more colours than it can grey levels, so colour can be used very

effectively to enhance small grey-level differences in a B/W image. The technique to display a monochrome image as a colour image is called *pseudo colour display*. A pseudo colour image is generated by assigning each grey level to a unique colour (Figure 1.6). This can be done by interactive colour editing or by automatic transformation based on certain logic. A common approach is to assign a sequence of grey levels to colours of increasing spectral wavelength and intensity.

The advantage of pseudo colour display is also its disadvantage. When a digital image is displayed in grey scale, using its DNs in a monochromatic display, the sequential numerical relationship between different DNs is effectively presented. This crucial information is lost in a pseudo colour display



**Figure 1.6** (a) An image in grey-scale (B/W) display; (b) the same image in a pseudo colour display; and (c) the brightest DNs are highlighted in red on a grey-scale background

because the colours assigned to various grey levels are not quantitatively related in a numeric sequence. Indeed, the image in a pseudo colour display is an image of symbols; it is no longer a digital image! We can regard the grey-scale B/W display as a special case of pseudo colour display in which a sequential grey scale based on DN levels is used instead of a colour scheme. Often, we can use a combination of B/W and pseudo colour display to highlight important information in particular DN ranges in colours over a grey-scale background as shown in Figure 1.6c.

### 1.3 Some key points

In this chapter, we learnt what a digital image is and the elements comprising a digital image and we also learnt about B/W and colour displays of digital images. It is important to remember these key points:

- A digital image is a raster dataset or a 2D array of numbers.
- Our perception of colours is based on the tristimulus theory of human vision. Any colour is composed of three primary colours: red, green and blue.
- Using an RGB colour cube, a colour can be expressed as a vector of the weighted summation of red, green and blue components.

- In image processing, colours are used as a tool for image information visualization. From this viewpoint, the true colour display is a special case of the general false colour display.
- Pseudo colour display results in the loss of the numerical sequential relationship of the image DNs. It is therefore no longer a digital image; it is an image of symbols.

### Questions

- 1.1 What is a digital image and how is it composed?
- 1.2 What are the major advantages of digital images over traditional hard copy images?
- 1.3 Describe the tristimulus colour theory and principle of RGB additive colour composition.
- 1.4 Explain the relationship between primary colours and complementary colours using a diagram.
- 1.5 Illustrate the colour cube in a diagram. How is a colour composed of RGB components? Describe the definition of the grey line in the colour cube.
- 1.6 What is a false colour composite? Explain the principle of using colours as a tool to visualize spectral information of multi-spectral images.
- 1.7 How is a pseudo colour display generated? What are the merits and disadvantages of pseudo colour display?

# 2

## Point Operations (Contrast Enhancement)

Contrast enhancement, sometimes called radiometric enhancement or histogram modification, is the most basic but also the most effective technique for optimizing the image contrast and brightness for visualization or for highlighting information in particular DN ranges.

Let  $X$  represent a digital image and  $x_{ij}$  be the DN of any pixel in the image at line  $i$  and column  $j$ . Let  $Y$  represent the image derived from  $X$  by a function  $f$  and  $y_{ij}$  be the output value corresponding to  $x_{ij}$ . Then a contrast enhancement can be expressed in the general form

$$y_{ij} = f(x_{ij}). \quad (2.1)$$

This processing transforms a single input image  $X$  to a single output image  $Y$ , through a function  $f$ , in such a way that the DN of an output pixel  $y_{ij}$  depends on and only on the DN of the corresponding input pixel  $x_{ij}$ . This type of processing is called a *point operation*. Contrast enhancement is a point operation that modifies the image brightness and contrast but does not alter the image size.

### 2.1 Histogram modification and lookup table

Let  $x$  represent a DN level of an image  $X$ ; then the number of pixels of each DN level  $h_i(x)$  is called the

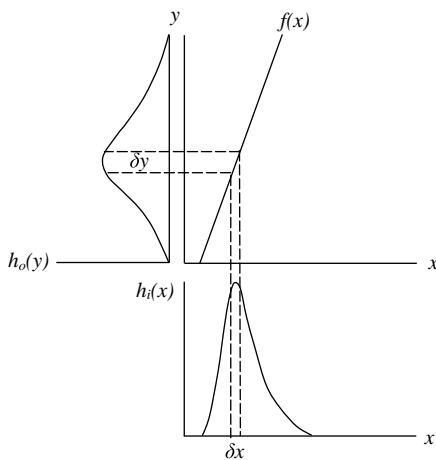
*histogram* of the image  $X$ . The  $h_i(x)$  can also be expressed as a percentage of the pixel number of a DN level  $x$  against the total number of pixels in the image  $X$ . In this case, in statistical terms,  $h_i(x)$  is a *probability density function*.

A histogram is a good presentation of the contrast, brightness and data distribution of an image. Every image has a unique histogram but the reverse is not necessarily true because a histogram does not contain any spatial information. As a simple example, imagine how many different patterns you can form on a  $10 \times 10$  grid chessboard using 50 white pieces and 50 black pieces. All these patterns have the same histogram!

It is reasonable to call a point operation a *histogram modification* because the operation only alters the histogram of an image but not the spatial relationship of image pixels. In Equation (2.1), point operation is supposed to be performed pixel by pixel. For the pixels with the same input DN but different locations ( $x_{ij} = x_{kl}$ ), the function  $f$  will produce the same output DN ( $y_{ij} = y_{kl}$ ). Thus the point operation is independent of pixel position. The point operation on individual pixels is the same as that on DN levels:

$$y = f(x). \quad (2.2)$$

As shown in Figure 2.1, suppose  $h_i(x)$  is a continuous function; as a point operation does not



**Figure 2.1** The principles of the point operation by histogram modification

change the image size, the number of pixels in the DN range  $\delta x$  in the input image  $X$  should be equal to the number of pixels in the DN range  $\delta y$  in the output image  $Y$ . Thus we have

$$h_i(x)\delta x = h_o(y)\delta y. \quad (2.3)$$

Let  $\delta x \rightarrow 0$ ; then  $\delta y \rightarrow 0$  and

$$h_i(x)dx = h_o(y)dy. \quad (2.4)$$

Therefore,

$$h_o(y) = h_i(x) \frac{dx}{dy} = h_i(x) \frac{dx}{f'(x)dx} = \frac{h_i(x)}{f'(x)}. \quad (2.5)$$

We can also write (2.5) as

$$h_o(y) = \frac{h_i(x)}{y'}.$$

The formula (2.5) shows that the histogram of the output image can be derived from the histogram of the input image divided by the first derivative of the point operation function.

For instance, given a linear function  $y = 2x - 6$ , then  $y' = 2$  and from (2.5) we have

$$h_o(y) = \frac{1}{2}h_i(x).$$

This linear function will produce an output image with a flattened histogram twice as wide and half as high as that of the input image and with all the DNs shifted to the left by three DN levels. This linear function stretches the image DN range to increase its contrast.

As  $f'(x)$  is the gradient of the point operation function  $f(x)$ , formula (2.5) thus indicates:

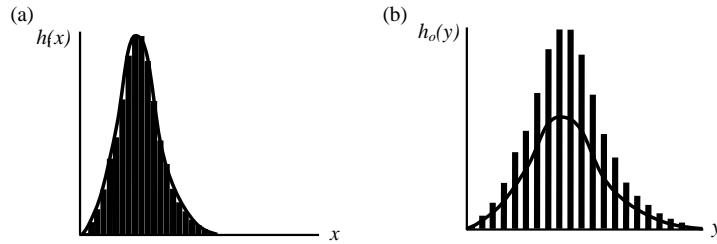
- (a) when the gradient of a point operation function is greater than 1, it is a stretching function which increases the image contrast;
- (b) when the gradient of a point operation function is less than 1, it is a compression function which decreases the image contrast;
- (c) if the gradient of a point operation function is negative, then the image becomes negative with black and white inverted.

For a nonlinear point operation function, this stretches and compresses different sections of DN levels, depending on its gradient at different DN levels, as shown later in the discussion on logarithmic and exponential point operation functions.

In the real case of an integer digital image, both  $h_i(x)$  and  $h_o(y)$  are discrete functions. Given a point operation  $y = f(x)$ , the DN level  $x$  in the image  $X$  is converted to a DN level  $y$  in output image  $Y$  and the number of pixels with DN value  $x$  in  $X$  is equal to that of pixels with DN value  $y$  in  $Y$ . Thus,

$$h_i(x) = h_o(y). \quad (2.6)$$

Equation (2.6) seems contradictory to Equation (2.3): that is,  $h_i(x)\delta x = h_o(y)\delta y$  for the case of a continuous function. In fact, Equation (2.6) is a special case of Equation (2.3) for  $\delta x = \delta y = 1$ , where 1 is the minimal DN interval for an integer digital image. Actually the point operation modifies the histogram of a digital image by moving the ‘histogram bar’ of each DN level  $x$  to a new DN level  $y$  according to the function  $f$ . The length of each histogram bar is not changed by the processing and thus no information is lost, but the distances between histogram bars are changed. For the given example above, the distance between histogram bars is doubled and thus the equivalent histogram averaged by the gap is flatter than the histogram of the input image (Figure 2.2). In this sense, Equation (2.3) always holds while Equation (2.6) is true only for individual histogram bars but not for the equivalent histogram. A point operation may merge several DN levels of an input image into one DN level of the output image. Equation (2.6) is then no longer true for some histogram bars and the operation results in information loss.



**Figure 2.2** Histograms before (a) and after (b) linear stretch for integer image data. Though the histogram bars in the histogram of the stretched image on the right are the same height as those in the original histogram on the left, the equivalent histogram drawn in the curve is wider and flatter because of the wider interval of these histogram bars

As point operation is in fact a histogram modification, it can be performed more efficiently using a *lookup table (LUT)*. An LUT is composed of DN levels of an input image  $X$  and their corresponding DN levels in the output image  $Y$ ; an example is shown in Table 2.1. When applying a point operation function to enhance an image, firstly the LUT is generated by applying the function  $y = f(x)$  to every DN level  $x$  of the input image  $X$  to generate the corresponding DN level  $y$  in the output image  $Y$ . Then, the output image  $Y$  is produced by just replacing  $x$  with its corresponding  $y$  for each pixel. In this case for an 8 bit image,  $y = f(x)$  needs to be calculated for no more than 256 times. If a point operation is performed without using an LUT,  $y = f(x)$  needs to be calculated as many times as the total number of pixels in the image. For a large image, the LUT approach speeds up processing dramatically, especially when the point operation function  $y = f(x)$  is a complicated one.

As most display systems can only display 8 bit integers in 0–255 grey levels, it is important to configure a point operation function in such a way that the value range of an output image  $Y$  is within 0–255.

## 2.2 Linear contrast enhancement

The point operation function for linear contrast enhancement (LCE) is defined as

$$y = ax + b. \quad (2.7)$$

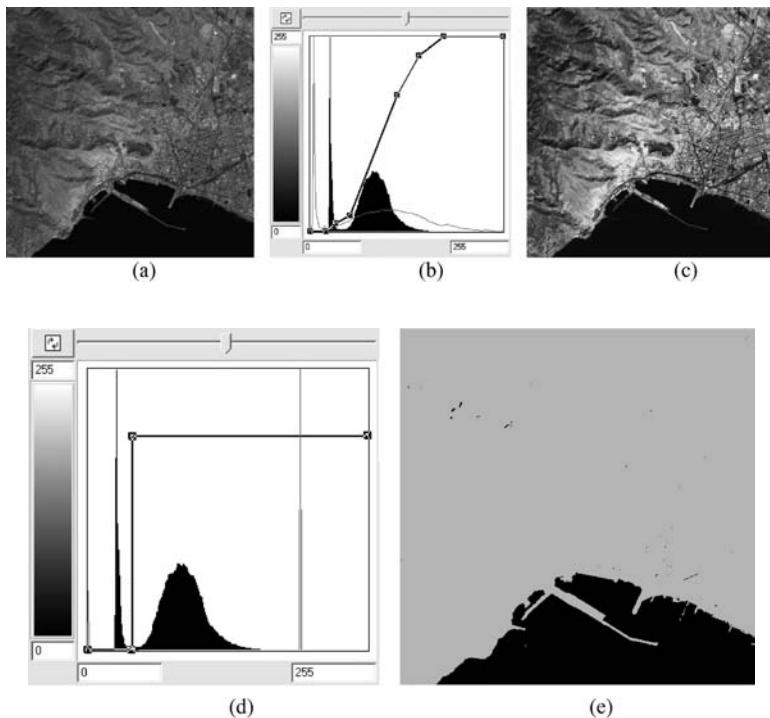
It is the simplest and one of the most effective contrast enhancement techniques. In this function, coefficient  $a$  controls the contrast of output images and  $b$  modifies the overall brightness by shifting the zero position of the histogram of  $y$  to  $-b/a$  (to the left if negative and to the right if positive). LCE improves image contrast without distorting the image information if the output DN range is wider than the input DN range. In this case, the LCE does nothing more than widen the increment of DN levels and shift histogram position along the image DN axis. For instance, the LCE function  $y = 2x - 6$  shifts the histogram  $h_i(x)$  to the left by three DN levels and doubles the DN increment of  $x$  to produce an output image  $Y$  with a histogram  $h_o(y) = h_i(x)/2$  that is two times wider than but half the height of the original.

There are several popular LCE algorithms available in most image processing software packages:

1. *Interactive linear stretch*: This changes  $a$  and  $b$  of formula (2.7) interactively to optimize the contrast and brightness of the output image based on the user's visual judgement.

**Table 2.1** An example LUT for a linear point operation function  $y = 2x - 6$

$x$	$y$
3	0
4	2
5	4
6	6
7	8
8	10
...	...
130	254



**Figure 2.3** Interactive PLS function for contrast enhancement and thresholding: (a) the original image; (b) the PLS function for contrast enhancement; (c) the enhanced image; (d) the PLS function for thresholding; and (e) the binary image produced by thresholding

2. *Piecewise linear stretch*: This uses several different linear functions to stretch different DN ranges of an input image (Figure 2.3a–c). Piecewise linear stretch (PLS) is a very versatile point operation function: it can be used to simulate a nonlinear function that cannot be easily defined by a mathematical function. Most image processing software packages have interactive PLS functionality allowing users to configure PLS for optimized visualization. Thresholding can be regarded as a special case of PLS as shown in Figure 2.3d–e, though in concept it is a conditional logic operation.
3. *Linear scale*: This automatically scales the DN range of an image to the full dynamic range of the display system (8 bits) based on the maximum and minimum of the input image  $X$ .

$$y = 255[x - \min(x)] / [\max(x) - \min(x)]. \quad (2.8)$$

In many modern image processing software packages, this function is largely redundant as the operation specified in (2.8) can be easily done

using an interactive PLS. However, formula (2.8) helps us to understand the principle.

4. *Mean/standard deviation adjustment*: This linearly stretches an image to make it satisfy a given mean ( $E_o$ ) and standard deviation ( $SD_o$ ):

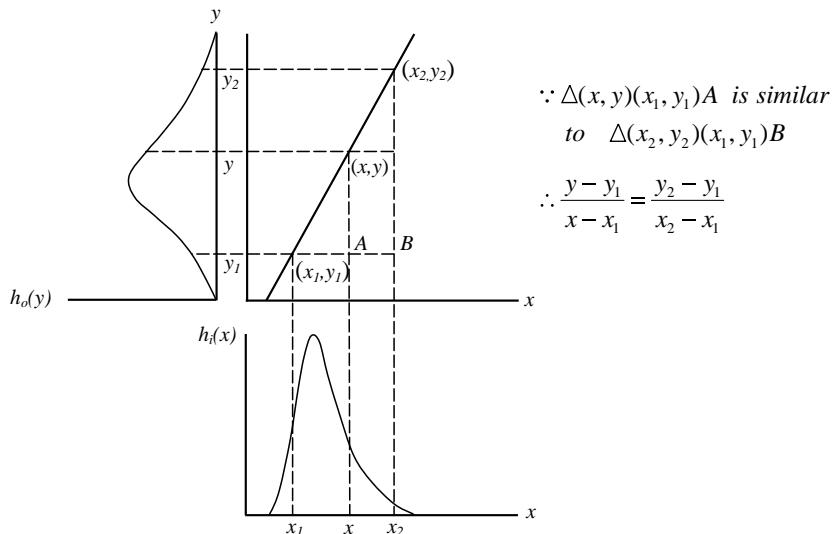
$$y = E_o + SD_o \frac{x - E_i}{SD_i} \quad \text{or} \quad y = \frac{SD_o}{SD_i}x + E_o - \frac{SD_o}{SD_i}E_i \quad (2.9)$$

where  $E_i$  and  $SD_i$  are the mean and standard deviation of the input image  $X$ .

These last two linear stretch functions are often used for automatic processing while, for interactive processing, PLS is the obvious choice.

### 2.2.1 Derivation of a linear function from two points

As shown in Figure 2.4, a linear function  $y = ax + b$  can be uniquely defined by two points  $(x_1, y_1)$  and



**Figure 2.4** Derivation of a linear function from two points of input image  $X$  and output image  $Y$

$(x_2, y_2)$  based on the formula

$$\frac{y - y_1}{x - x_1} = \frac{y_2 - y_1}{x_2 - x_1}.$$

Given  $x_1 = \min(x)$ ,  $x_2 = \max(x)$  and  $y_1 = 0$ ,  $y_2 = 255$ , we then have

$$\frac{y}{x - \min(x)} = \frac{255}{\max(x) - \min(x)}.$$

Thus  $y = 255(x - \min(x))/(\max(x) - \min(x))$ .

Similarly, linear functions for mean and standard deviation adjustment defined in (2.9) can be derived from either

$$x_1 = E_i, \quad x_2 = E_i + SD_i, \quad y_1 = E_o, \quad y_2 = E_o + SD_o$$

or

$$x_1 = E_i, \quad x_2 = E_i - SD_i, \quad y_1 = E_o, \quad y_2 = E_o - SD_o.$$

## 2.3 Logarithmic and exponential contrast enhancement

Logarithmic and exponential functions are inverse operations of one another. For contrast enhancement, the two functions modify the image histograms in opposite ways. Both logarithmic and exponential functions change the shapes of image

histograms and distort the information in original images.

### 2.3.1 Logarithmic contrast enhancement

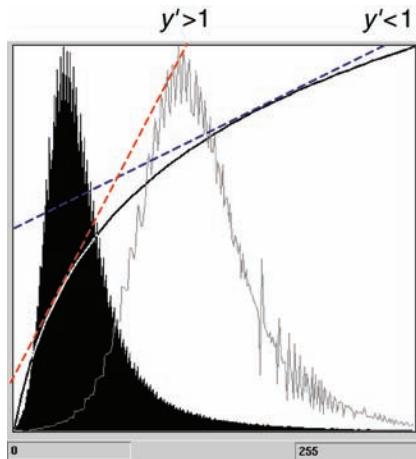
The general form of the logarithmic function used for image processing is defined as

$$y = b \ln(ax + 1). \quad (2.10)$$

Here  $a (> 0)$  controls the curvature of the logarithmic function while  $b$  is a scaling factor to make the output DNs fall within a given value range, and the shift 1 is to avoid the zero value at which the logarithmic function loses its meaning. As shown in Figure 2.5, the gradient of the function is greater than 1 in the low DN range, thus it spreads out low DN values, while in the high DN range the gradient of the function is less than 1 and so compresses high DN values. As a result, logarithmic contrast enhancement shifts the peak of the image histogram to the right and highlights the details in dark areas in an input image. Many images have histograms similar in form to logarithmic normal distributions. In such cases, a logarithmic function will effectively modify the histogram to the shape of a normal distribution.

We can slightly modify formula (2.10) to introduce a shift constant  $c$ :

$$y = b \ln(ax + 1) + c. \quad (2.11)$$



**Figure 2.5** Logarithmic contrast enhancement function

This function allows the histogram of the output image to shift by  $c$ .

### 2.3.2 Exponential contrast enhancement

The general form of the exponential function used for image processing is defined as

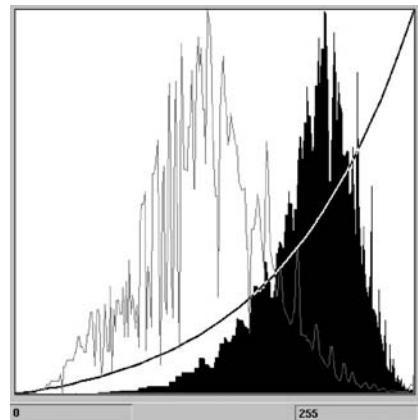
$$y = b e^{ax+1}. \quad (2.12)$$

Here again,  $a (>0)$  controls the curvature of the exponential function while  $b$  is a scaling factor to make the output DNs fall within a given value range, and the exponential shift 1 is to avoid the zero value because  $e^0 \equiv 1$ . As the inverse of the logarithmic function, exponential contrast enhancement shifts the image histogram to the left by spreading out high DN values and compressing low DN values to enhance detail in light areas at the cost of suppressing the tone variation in the dark areas (Figure 2.6). Again, we can introduce a shift parameter  $c$ , to modify the exponential contrast enhancement function as below:

$$y = b e^{ax+1} + c. \quad (2.13)$$

## 2.4 Histogram equalization

Histogram equalization (HE) is a very useful contrast enhancement technique. It transforms an



**Figure 2.6** Exponential contrast enhancement function

input image to an output image with a uniform (equalized) histogram. The key point of HE is to find the function that converts  $h_i(x)$  to  $h_o(y) = A$ , where  $A$  is a constant. Suppose image  $X$  has  $N$  pixels and the desired output DN range is  $L$  (the number of DN levels). Then

$$h_o(y) = A = \frac{N}{L}. \quad (2.14)$$

According to (2.4)

$$dy = h_i(x)dx/h_o(y) = \frac{L}{N}h_i(x)dx. \quad (2.15)$$

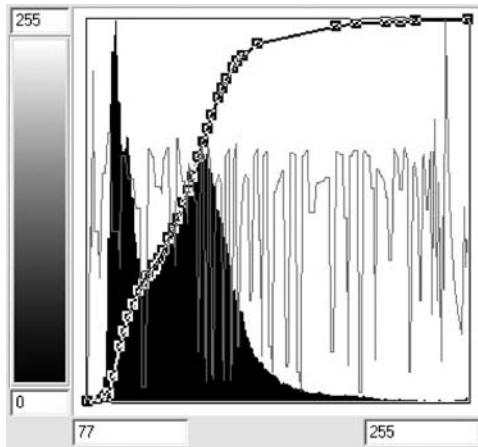
Thus, the HE function is

$$y = \frac{L}{N} \int h_i(x)dx = \frac{L}{N}H_i(x). \quad (2.16)$$

As the histogram  $h_i(x)$  is essentially the probability density function of  $X$ , the  $H_i(x)$  is the *cumulative distribution function* of  $X$ . The calculation of  $H_i(x)$  is simple for a discrete function in the case of digital images. For a given DN level  $x$ ,  $H_i(x)$  is equal to the total number of those pixels with DN values no greater than  $x$ :

$$H_i(x) = \sum_{k=0}^x h_i(k). \quad (2.17)$$

Theoretically, HE can be achieved if  $H_i(x)$  is a continuous function. However, as  $H_i(x)$  is a discrete function for an integer digital image, HE can only produce a relatively flat histogram mathematically



**Figure 2.7** Histogram of histogram equalization

equivalent to an equalized histogram, in which the distance between histogram bars is proportional to their heights (Figure 2.7).

The idea behind the HE contrast enhancement is that the data presentation of an image should be evenly distributed across the whole value range. In reality, however, HE often produces images with too high a contrast. This is because natural scenes are more likely to follow normal (Gaussian) distributions and, consequently, the human eye is adapted to be more sensitive for discriminating subtle grey-level changes, of intermediate brightness, than of very high and very low brightness.

## 2.5 Histogram matching and Gaussian stretch

Histogram matching (HM) is a point operation that transforms an input image to make its histogram

match a given shape defined by either a mathematical function or the histogram of another image. It is particularly useful for image comparison and differencing. If the two images in question are modified to have similar histograms, the comparison will be on a fair basis.

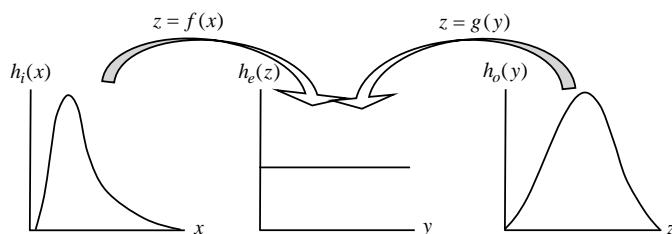
HM can be implemented by applying HE twice. Formula (2.14) implies that an equalized histogram is only decided by image size  $N$  and the output DN range  $L$ . Images of the same size always have the same equalized histogram for a fixed output DN range and thus HE can act as a bridge to link images of the same size but different histograms (Figure 2.8). Consider  $h_i(x)$  as the histogram of an input image and  $h_o(y)$  the reference histogram to be matched. Suppose  $z = f(x)$  is the HE function to transform  $h_i(x)$  to an equalized histogram  $h_e(z)$ , and  $z = g(y)$  the HE function to transform the reference histogram  $h_o(y)$  to the same equalized histogram  $h_e(z)$ . Then

$$z = g(y) = f(x).$$

Thus

$$y = g^{-1}(z) = g^{-1}\{f(x)\}. \quad (2.18)$$

Recall from formula (2.16) that  $f(x)$  and  $g(y)$  are the cumulative distribution functions of  $h_i(x)$  and  $h_o(y)$  individually. Thus HM can be easily implemented by a three-column LUT containing corresponding DN levels of  $x$ ,  $z$  and  $y$ . An input DN level  $x$  will be transformed to an output DN level  $y$  sharing the same  $z$  value. As shown in Table 2.2, for  $x = 5$ ,  $z = 3$ , while for  $y = 0$ ,  $z = 3$ . Thus for an input  $x = 5$ , the LUT converts to an output  $y = 0$  and so on. The output image  $Y$  will have a histogram that matches the reference histogram  $h_o(y)$ .



**Figure 2.8** Histogram equalization acts as a bridge for histogram matching

**Table 2.2** An example LUT for histogram matching

<i>x</i>	<i>z</i>	<i>y</i>
5	3	0
6	4	2
7	5	4
8	6	5
...	...	...

If the reference histogram  $h_o(y)$  is defined by a Gaussian distribution function

$$h_o(y) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(y - \bar{x})^2}{2\sigma^2}\right) \quad (2.19)$$

where  $\sigma$  is the standard deviation and  $\bar{x}$  the mean of image  $X$ , the HM transformation is then called *Gaussian stretch* since the resultant image has a histogram in the shape of a Gaussian distribution.

## 2.6 Balance contrast enhancement technique

Colour bias is one of the main causes of poor colour composite images. For RGB colour composition, if the average brightness of one image band is significantly higher or lower than the other two, the composite image will show obvious colour bias. To eliminate this, the three bands used for colour composition must have an equal value range and mean. The balance contrast enhancement technique (BCET) is a simple solution to this problem. Using a parabolic function derived from an input image, BCET can stretch (or compress) the image to a given value range and mean without changing the basic shape of the image histogram. Thus three image bands for colour composition can be adjusted to the same value range and mean to achieve a balanced colour composite.

The BCET based on a parabolic function is

$$y = a(x - b)^2 + c. \quad (2.20)$$

This general form of parabolic function is defined by three coefficients:  $a$ ,  $b$  and  $c$ . It is therefore capable of adjusting three image parameters:

minimum, maximum and mean. The coefficients  $a$ ,  $b$  and  $c$  can be derived based on the minimum, maximum and mean ( $l$ ,  $h$  and  $e$ ) of the input image  $X$  and the given minimum, maximum and mean ( $L$ ,  $H$  and  $E$ ) for the output image  $Y$  as follows:

$$\begin{aligned} b &= \frac{h^2(E-L) - s(H-L) + l^2(H-E)}{2[h(E-L) - e(H-L) + l(H-E)]} \\ a &= \frac{H-L}{(h-l)(h+l-2b)} \\ c &= L - a(l-b)^2 \end{aligned} \quad (2.21)$$

where  $s$  is the mean square sum of input image  $X$ ,

$$s = \frac{1}{N} \sum_{i=1}^N x_i^2.$$

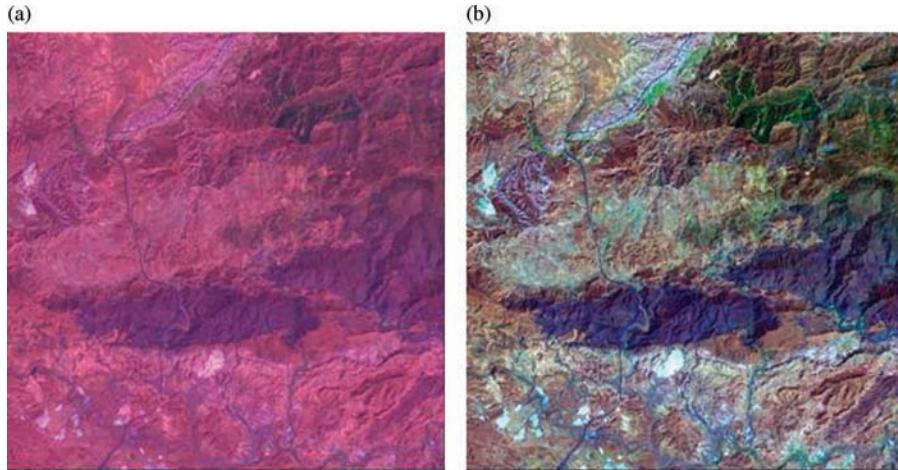
Figure 2.9 illustrates a comparison between RGB colour composites using the original band 5, 4 and 1 of an ETM+ sub-scene and the same bands after BCET stretch. The colour composite of the original bands (Figure 2.9a) shows strong colour bias to magenta as the result of much lower brightness in band 4, displayed in green. This colour bias is completely removed by BCET which stretches all the bands to the same value range 0–255 and mean 110 (Figure 2.9b). The BCET colour composite in Figure 2.9b presents various terrain materials (rock types, vegetation, etc.) in much more distinctive colours than those in the colour composite of the original image bands in Figure 2.9a. An interactive PLS may achieve similar results but without quantitative control.

### 2.6.1 \*Derivation of coefficients, $a$ , $b$ and $c$ for a BCET parabolic function (Liu, 1991)

Let  $x_i$  represent any pixel of an input image  $X$ , with  $N$  pixels. Then the minimum, maximum and mean of  $X$  are

$$\begin{aligned} l &= \min(x_i), \quad h = \max(x_i), \\ e &= \frac{1}{N} \sum_{i=1}^N x_i, \quad i = 1, 2, \dots, N. \end{aligned}$$

Suppose  $L$ ,  $H$  and  $E$  are the desired minimum, maximum and mean for the output image  $Y$ . Then



**Figure 2.9** Colour composites of ETM+ bands 5, 4 and 1 in red, green and blue: (a) colour composite of the original bands showing magenta cast as the result of colour bias; and (b) BCET colour composite stretching all the bands to an equal value range of 0–255 and mean of 110

we can establish following equations:

$$\begin{aligned} L &= a(l-b)^2 + c \\ H &= a(h-b)^2 + c \\ E &= \frac{1}{N} \sum_{i=1}^N [a(x_i-b)^2 + c]. \end{aligned} \quad (2.22)$$

Solving for  $b$  from (2.22),

$$b = \frac{h^2(E-L) - s(H-L) + l^2(H-E)}{2[h(E-L) - e(H-L) + l(H-E)]} \quad (2.23)$$

where

$$s = \frac{1}{N} \sum_{i=1}^N x_i^2.$$

With  $b$  known,  $a$  and  $c$  can then be resolved from (2.22) as

$$a = \frac{H-L}{(h-l)(h+l-2b)} \quad (2.24)$$

$$c = L - a(l-b)^2. \quad (2.25)$$

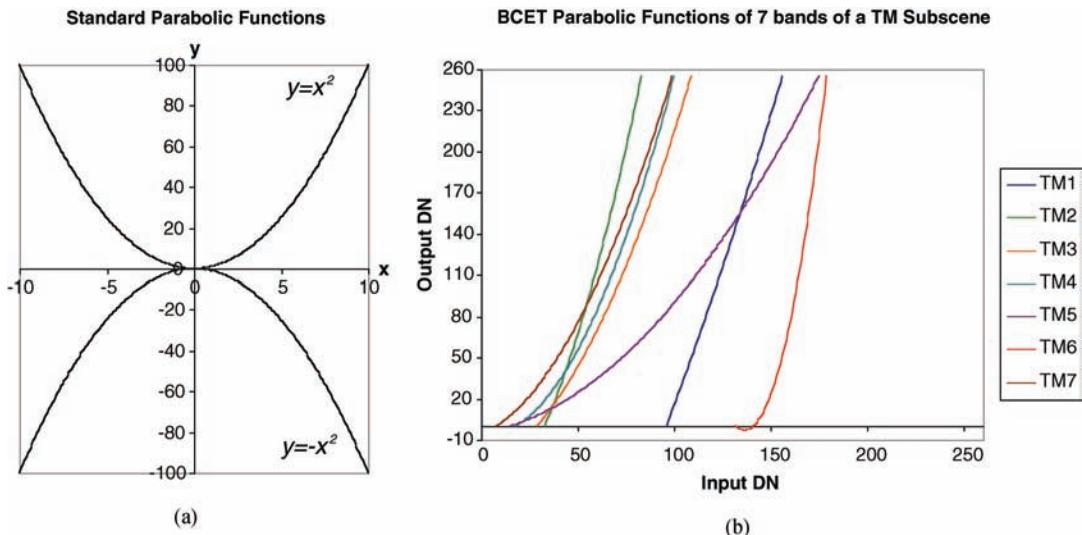
The parabolic function is an even function (Figure 2.10a). Coefficients  $b$  and  $c$  are the coordinates of the turning point of the parabola which determine the section of the parabola to be utilized by the BCET function. In order to perform BCET, the turning point and its nearby section of the

parabola should be avoided, so that only the section of the monotonically increasing branch of the curve is used. This is possible for most cases of image contrast enhancement.

From the solutions of  $a$ ,  $b$  and  $c$  in Equations (2.23)–(2.25), we can make the following observations:

- (a) If  $b < l$  and  $a > 0$ , the parabola is open upwards and a section of the right (monotonically increasing) branch of the parabola is used in BCET.
- (b) If  $b > h$  and  $a < 0$ , the parabola is open downwards and a section of the left (monotonically increasing) branch of the parabola is used in BCET.
- (c) If  $l < b < h$ , then BCET fails to avoid the turning point of the parabola and malfunctions.

For example, Table 2.3 shows the minimum ( $l$ ) maximum ( $h$ ) and mean ( $e$ ) of seven band images of a Landsat TM sub-scene and the corresponding coefficients of the BCET parabolic functions. Using these parabolic functions, images of bands 1–5 and 7 are all successfully stretched to the given value range and mean:  $L = 0$ ,  $H = 255$  and  $E = 100$  as shown in the right part of the table. The only exception is the band 6 image because  $l < b < h$  and BCET malfunctions. As illustrated in Figure 2.10b, the BCET



**Figure 2.10** (a) Standard parabolas  $y = x^2$  and  $y = -x^2$ , the cases of  $a = \pm 1, b = 0, c = 0$  for  $y = a(x - b)^2 + c$ ; and (b) BCET parabolic functions for seven band images of a TM sub-scene. The parabola for the band 6 image in red involves the turning point and both branches and is therefore not usable

parabolic function for band 6 involves the turning point and both branches of the parabola within the value range of this image, unlike all the other bands where only one monotonic branch is used.

## 2.7 Clipping in contrast enhancement

In digital images, a few pixels (often representing noise) may occupy a wide value range at the low and high ends of histograms. In such cases, setting a proper cut-off to clip both ends of the histogram in contrast enhancement is necessary to make effective use of the dynamic range of a display device. Clipping is often given as a percentage of the total number of pixels in an image. For instance, if 1% and 99% are set as the cut-off limits for the low and high ends of the histogram of an image, the image is then stretched to set the DN level  $x_l$ , where  $H_i(x_l) = 1\%$ , to 0 and DN level  $x_h$ , where  $H_i(x_h) = 99\%$ , to 255 for an 8 bit per pixel per channel display in the output image.

This simple treatment often improves image display quality significantly, especially when the image looks hazy because of atmospheric scattering. When using BCET, the input minimum ( $l$ ) and

maximum ( $h$ ) should be determined based on appropriate cut-off levels of  $x_l$  and  $x_h$ .

## 2.8 Tips for interactive contrast enhancement

The general purpose of contrast enhancement is to optimize visualization. Often after quite complicated image processing, you will need to apply interactive contrast enhancement to view the results properly. After all, you need to be able to see the image! Visual observation is always the most effective way to judge image quality. This does not sound technical enough for digital image processing but this golden rule is quite true! On the other hand, the histogram gives you a quantitative description of image data distribution and so can also effectively guide you to improve the image visual quality. As mentioned earlier, the business of contrast enhancement is histogram modification and so you should find the following guidelines useful:

1. Make full use of the dynamic range of the display system. This can be done by specifying the actual limits of the input image to be displayed

**Table 2.3** Derivation of BCET parabolic functions for the seven band images of a Landsat TM sub-scene to stretch each band to  $L = 0$ ,  $H = 255$  and  $E = 255$ 

Image bands	Input image			BCET coefficients			BCET output image		
	Min ( $l$ )	Max ( $h$ )	Mean ( $e$ )	$a$	$b$	$c$	Min ( $L$ )	Max ( $H$ )	Mean ( $E$ )
TM1	96	156	120.78	0.00857	-121.91	-407.03	0	255	100.00
TM2	33	83	55.47	0.03594	-12.95	-75.88	0	255	99.90
TM3	29	109	66.47	0.01934	-13.4	-34.78	0	255	99.99
TM4	18	100	59	0.02776	2.97	-6.26	0	255	100.05
TM5	15	175	94.33	0.00718	-15.97	-6.89	0	255	99.96
TM6	132	179	157.88	0.14302	136.53	-2.94	Malfunction, not used		
TM7	8	99	52.63	0.02072	-14.11	-10.13	0	255	100.02

in 0 and 255 for an 8 bit display. Here percentage clipping is useful to avoid large gaps in either end of the histogram.

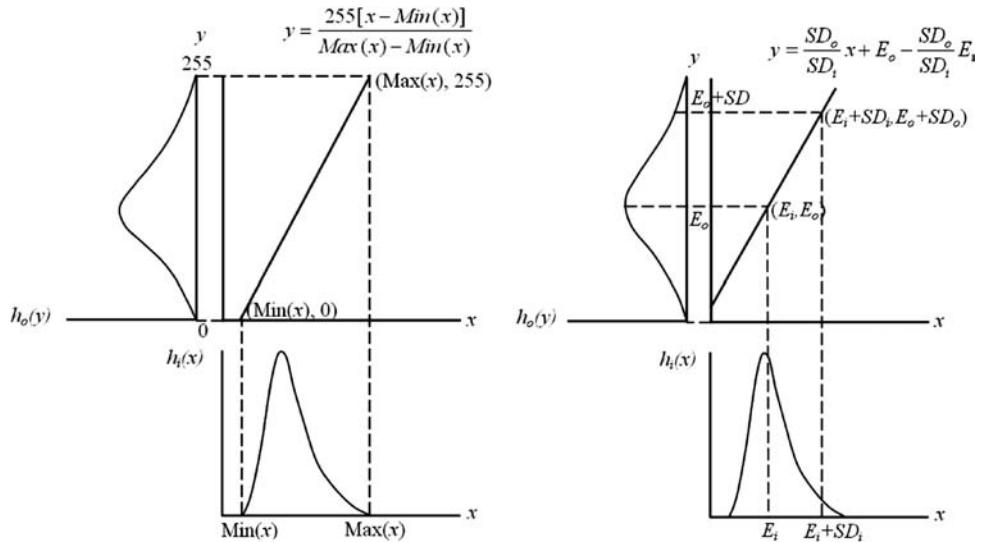
2. Adjust the histogram to put its peak near the centre of the display range. For many images, the peak may be slightly skewed towards the left to achieve the best visualization, unless the image is dominated by bright features in which case the peak could skew to the right.
3. Note that, as implied by formula (2.5), a point operation function modifies an image histogram according to the function's gradient or slope  $f'(x)$ :
  - (a) If gradient = 1 (slope =  $45^\circ$ ), the function does nothing and the image is not changed.
  - (b) If gradient > 1 (slope >  $45^\circ$ ), the function stretches the histogram to increase image contrast.
  - (c) If gradient < 1 (slope <  $45^\circ$ ) and non-negative, the function compresses the histogram to decrease image contrast.

A common approach in the PLS is therefore to use functions with slope  $>45^\circ$  to spread the peak section and those with slope  $<45^\circ$  to compress the tails at both ends of the histogram.

## Questions

- 2.1 What is a point operation in image processing? Give the mathematical definition.
- 2.2 Using a diagram, explain why a point operation is also called histogram modification.
- 2.3 Given the following point operation functions, derive the output histograms  $h_o(y)$  from the input histogram  $h_i(x)$ :
 
$$y = 3x - 8; \quad y = 2.5x^2 - 3x + 2; \quad y = \sin(x).$$
- 2.4 Try to derive the *linear scale* functions and the *mean and standard deviation adjustment* functions defined by formulae (2.8) and (2.9). (See the answer at the end in Figure 2.11.)
- 2.5 Given Figure 2.6 of exponential contrast enhancement, roughly mark the section of the exponential function that stretches the input image and the section that compresses the input image and explain why (refer to Figure 2.5).

- 2.6 How is histogram equalization (HE) achieved?  
How is HE used to achieve histogram matching?
- 2.7 What type of function does a BCET use and how  
is balanced contrast enhancement achieved?
- 2.8 Try to derive the coefficients  $a$ ,  $b$  and  $c$  in the  
BCET function  $y = a(x - b)^2 + c$ .
- 2.9 What is clipping and why is it often essential for  
image display?



**Figure 2.11** Derivation of the linear stretch function and mean/standard deviation adjustment function

# 3

## Algebraic Operations (Multi-image Point Operations)

For multi-spectral or, more generally, multi-layer images, algebraic operations such as the four basic arithmetic operations ( $+$ ,  $-$ ,  $\times$ ,  $\div$ ), logarithms, exponentials and trigonometric functions can be applied to the DNs of different bands for each pixel to produce a new image. Such processing is called image algebraic operation. Algebraic operations are performed pixel by pixel among DNs of spectral bands (or layers) for each pixel without involving neighbourhood pixels. They can therefore be considered as *multi-image point operations* defined as follows:

$$y = f(x_1, x_2, \dots, x_n) \quad (3.1)$$

where  $n$  is the number of bands or layers.

Obviously, all the images involving algebraic operations should be precisely co-registered.

To start with, let us consider the four basic arithmetic operations: addition, subtraction, multiplication and division. In multi-image point operations, arithmetic processing is sometimes the same as matrix operations, such as addition and subtraction, but sometimes totally different from and much simpler than matrix operations, such as image multiplication and division. As the image algebraic operation is entirely local, that is pixel-to-pixel based, we can generalize the description. Let  $X_i$ ,

$i = 1, 2, \dots, n$ , represent both the  $i$ th band image and any pixel in the  $i$ th band image of an  $n$ -band imagery dataset  $\mathbf{X}$ ,  $X_i \in \mathbf{X}$ , and  $Y$  the output image as well as any pixel in the output image.

### 3.1 Image addition

This operation produces a weighted summation of two or more images:

$$Y = \frac{1}{k} \sum_{i=1}^n w_i X_i \quad (3.2)$$

where  $w_i$  is the weight of image  $X_i$  and  $k$  is a scaling factor.

If  $w_i = 1$  for  $i = 1, \dots, n$  and  $k = n$ , formula (3.2) defines an average image.

An important application of image addition is to reduce noise and increase the signal to noise ratio (SNR). Suppose each image band of an  $n$ -band multi-spectral image is contaminated by an additive noise source  $N_i$  ( $i = 1, 2, \dots, n$ ); then the noise pixels are not likely to occur at the same positions in different bands and thus a noise pixel DN in band  $i$  will be averaged with the non-noise DNs in the other  $n - 1$  bands. As a result the noise will be

largely suppressed. It is proved from signal processing theory that, of  $n$  duplications of an image, each contaminated by the same level of random noise, the SNR of the sum image of these  $n$  duplications equals the square root  $n$  times the SNR of any individual duplication:

$$SNR_y = \sqrt{n} \cdot SNR_i. \quad (3.3)$$

The formula (3.3) implies that for an  $n$ -band multi-spectral image, the summation of all the bands can increase SNR by about  $\sqrt{n}$  times. For instance, if we average bands 1–4 of a Landsat TM image, the SNR of this average image is about two times ( $\sqrt{4} = 2$ ) of that of each individual band.

You may notice in our later chapters on topics of RGB–IHS (red, green and blue to intensity, hue and saturation) transformation and principal component analysis (PCA) that an intensity component derived from RGB–IHS transformation is an average image of the R, G and B component images and, in most cases, the first principal component is a weighted sum image of all the images involving PCA operations.

### 3.2 Image subtraction (differencing)

Image subtraction produces a difference image from two input images:

$$Y = \frac{1}{k} (w_i X_i - w_j X_j). \quad (3.4)$$

The weights  $w_i$  and  $w_j$  are important to ensure that balanced differencing is performed. If the brightness of  $X_i$  is significantly higher than that of  $X_j$ , for instance, the difference image  $X_i - X_j$  will be dominated by  $X_i$  and, as a result, the true difference between the two images will not be effectively revealed. To produce a ‘fair’ difference image, BCET or histogram matching (matching the histogram of  $X_i$  to that of  $X_j$ ) may be applied as a pre-processing step. Whichever method is chosen, the differencing that follows should then be performed with equal weighting ( $w_i = w_j = 1$ ).

Subtraction is one of the simplest and most effective techniques for selective spectral enhance-

ment and it is also useful for change detection and removal of background illumination bias. However, in general, subtraction reduces the image information and decreases image SNR because it removes the common features while retaining the random noise in both images.

Band differences of multi-spectral images are successfully used for studies of vegetation, land use and geology. As shown in Figure 3.1, the band difference of  $TM3 - TM1$  (R – B) highlights iron oxides;  $TM4 - TM3$  (NIR – Red) enhances vegetation; and  $TM5 - TM7$  is effective for detecting hydrated (clay) minerals (i.e. those containing the  $OH^-$  ion; refer to Table A.1 in Appendix A for the spectral wavelengths of Landsat TM). These three difference images can be combined to form an RGB colour composite image to highlight iron oxides, vegetation and clay minerals in red, green and blue as well as other ground objects in various colours. In many cases, subtraction can achieve similar results to division (ratio) and the operation is simpler and faster.

The image subtraction technique is also widely used for background noise removal in microscopic image analysis. An image of the background illumination field (as a reference) is captured before the target object is placed in the field. The second image is then taken with the target object in the field. The difference image between the two will retain the target while the effects of the illumination bias and background noise are cancelled out.

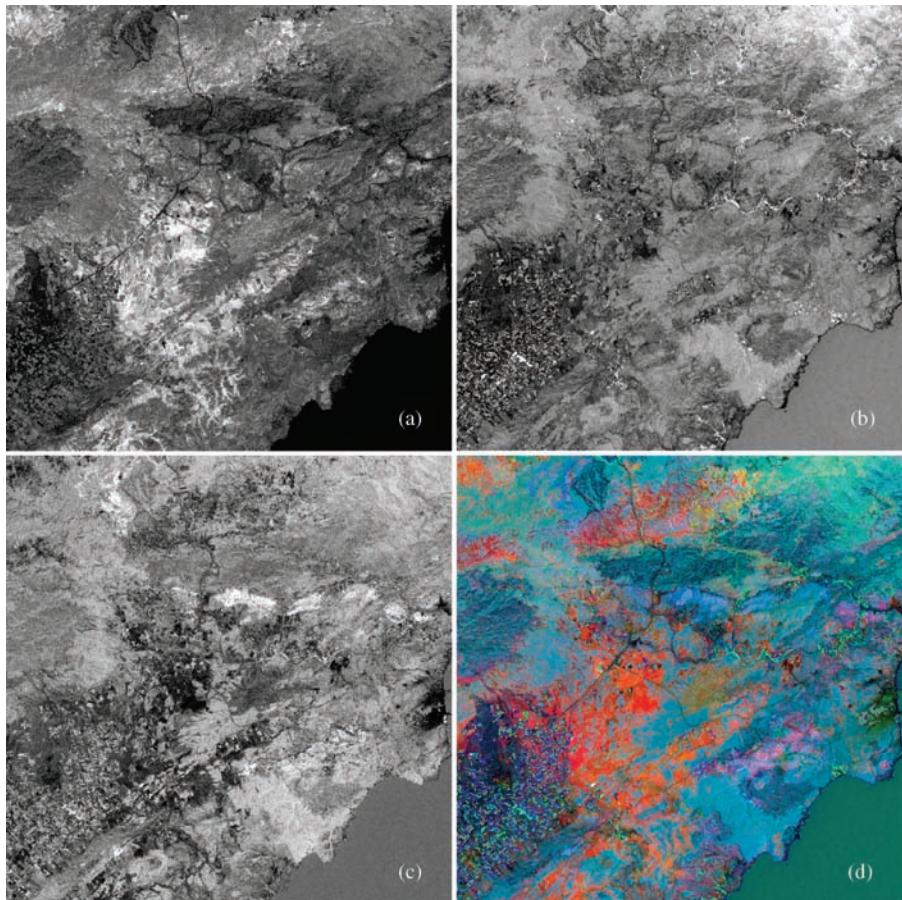
### 3.3 Image multiplication

Image multiplication is defined as

$$Y = X_i \cdot X_j. \quad (3.5)$$

Here the image multiplication is performed pixel by pixel; at each image pixel, its band  $i$  DN is multiplied with band  $j$  DN. This is fundamentally different from matrix multiplication. A digital image is a 2D array, but it is *not* a matrix.

A multiplication product image often has much greater DN range than the dynamic range of the display devices and thus need to be rescaled before display. Most image processing software packages



**Figure 3.1** Difference images of a Landsat TM image: (a)  $TM_3 - TM_1$  highlights red features often associated to iron oxides; (b)  $TM_4 - TM_3$  detects the diagnostic ‘red edge’ features of vegetation; (c)  $TM_5 - TM_7$  enhances the clay mineral absorption features in SWIR spectral range; and (d) the colour composite of  $TM_3 - TM_1$  in red,  $TM_4 - TM_3$  in green and  $TM_5 - TM_7$  in blue highlights iron oxide, vegetation and clay minerals in red, green and blue colours

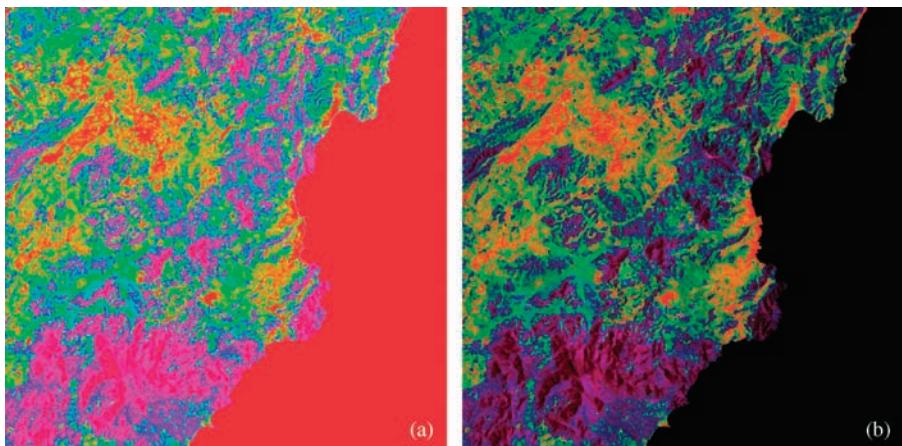
can display any image based on its actual value range which is then fitted into a 0–255 display range.

One application of multiplication is *masking*. For instance, if  $X_i$  is a mask image composed of DN values 0 and 1, the pixels in image  $X_j$  which correspond to 0 in  $X_i$  will become 0 (masked off) and others will remain unchanged in the product image  $Y$ . This operation could be achieved more efficiently using a logical operation of a given condition. Another application is image modulation. For instance, topographic features can be added back to a colour-coded classification image by using a panchromatic image (as an intensity component) to modulate the three colour compo-

nents (red, green and blue) of the classification image as follows:

1. Produce red ( $R$ ), green ( $G$ ) and blue ( $B$ ) component images from the colour-coded classification image.
2. Use the relevant panchromatic image ( $I$ ) to modulate the  $R$ ,  $G$  and  $B$  components:  $R \times I$ ,  $G \times I$  and  $B \times I$ .
3. Colour composition using  $R \times I$ ,  $G \times I$  and  $B \times I$ .

This process is, in some image processing software packages, automated by draping a colour-coded classification image on an intensity image layer (Figure 3.2).



**Figure 3.2** Multiplication for image modulation: (a) a colour-coded classification image; and (b) the intensity-modulated classification image

### 3.4 Image division (ratio)

Image division is a very popular technique, also known as an image *ratio*. The operation is defined as

$$Y = \frac{X_i}{X_j}. \quad (3.6)$$

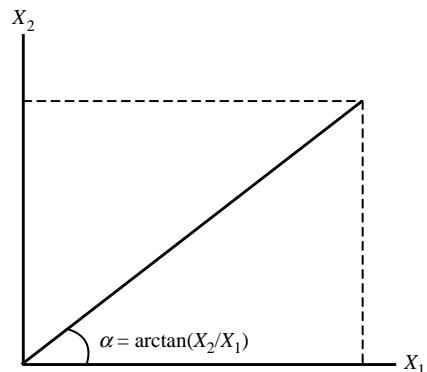
In order to carry out image division, certain protection is needed to avoid overflow, in case a number is divided by 0. A commonly used trick in this context is to change 0 to 1 whenever 0 becomes a divisor. A better approach is to shift the value range of the denominator image upwards, by 1, to avoid 0. For an 8 bit image, this shift changes the image DN range from 0–255 to 1–256 which just exceeds 8 bits. This was a problem in the older generation of image processing systems before the 1990s but is no longer in most modern image processing software packages where the image processing is performed based on the double-precision, floating-point data type.

A ratio image  $Y$  is an image of real numbers instead of integers. If both  $X_i$  and  $X_j$  are 8 bit images, the possible maximum value range of  $Y$  is 0, [1/255, 1], (1, 255]. Instead of a much simpler notation, [0, 255], we deliberately write the value range in such a way to emphasize that the value range [1/255, 1] may contain just as much information as that in the much wider value range (1, 255]! A popular approach for displaying a ratio image on an 8 bit per pixel per channel display system is to scale the image into a

0–255 DN range; many image processing software packages may perform the operation automatically. This might result in up to 50% information loss because the information recorded in value range [1/255, 1] could be compressed into a few DN levels.

If we consider an image ratio as a coordinate transformation from a Cartesian coordinate system to a polar coordinate system (Figure 3.3), rather than a division operation, then

$$\begin{aligned} Y &= \frac{X_i}{X_j} = \tan(\alpha) \\ \alpha &= \arctan\left(\frac{X_i}{X_j}\right). \end{aligned} \quad (3.7)$$



**Figure 3.3** Ratio as a coordinate transformation from a Cartesian coordinates system to a polar coordinates system

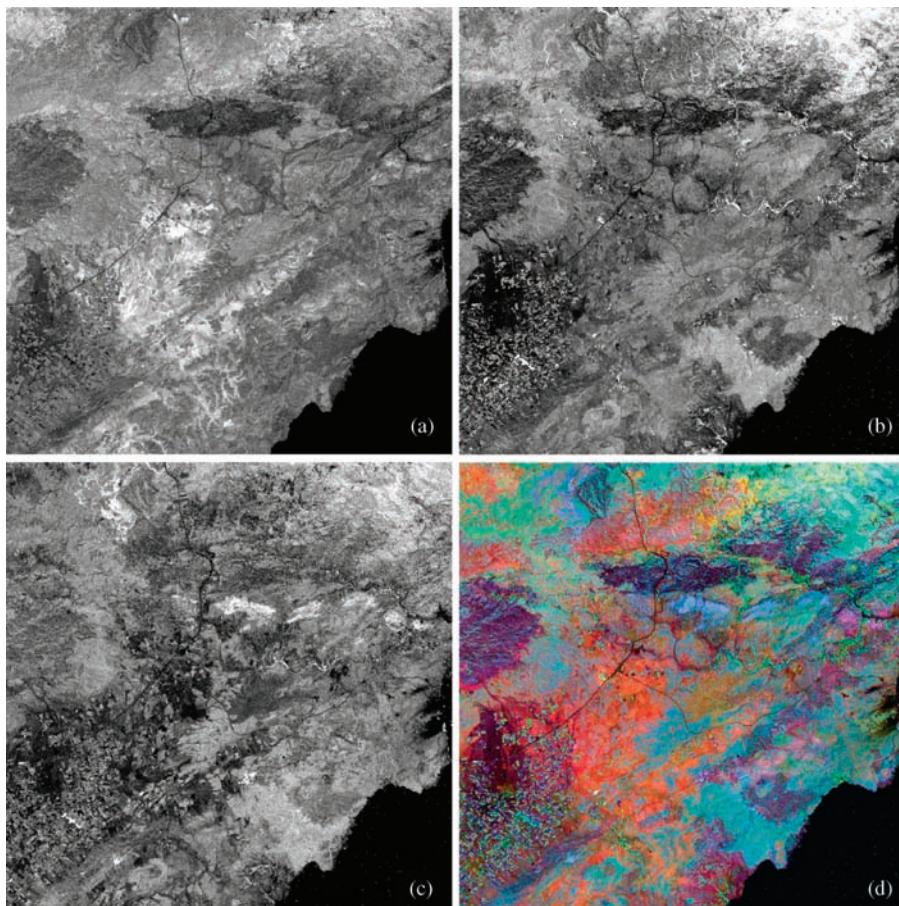
Ratio image  $Y$  is actually a tangent image of the angle  $\alpha$ . The information of a ratio image is evenly presented by angle  $\alpha$  in value range  $[0, \pi/2]$  instead of by  $Y = \tan(\alpha)$  in value range  $[0, 255]$ . Therefore, to achieve a ‘fair’ linear scale stretch of a ratio image, it is necessary to convert  $Y$  to  $\alpha$  by formula (3.7). A linear scale can then be performed as

$$\beta = 255 \frac{\alpha - \min(\alpha)}{\max(\alpha) - \min(\alpha)}. \quad (3.8)$$

After all, the above transform may not always be necessary. Ratios are usually designed to highlight the target features as high-ratio DNs. In this case, the direct stretch of ratio image  $Y$  may enhance the target features well but at the expense of the information represented by low-ratio DNs. From this

sense and as an example, it is important to notice that although ratios  $TM1/TM3$  and  $TM3/TM1$  are reciprocals of one another mathematically and so contain the same information, they are different in terms of digital image display! Remember: when you design a ratio, make sure the target information is highlighted by high values in the ratio image.

Ratio is an effective technique for selectively enhancing spectral features. Ratio images derived from different band pairs are often used to generate ratio colour composites in an RGB display. For instance, a colour composite of  $TM5/TM7$  (blue),  $TM4/TM3$  (green) and  $TM3/TM1$  (red) may highlight clay minerals in blue, vegetation in green and iron oxide in red (Figure 3.4). It is interesting to compare Figure 3.1d with Figure 3.4d to notice the



**Figure 3.4** Ratio images and ratio colour composite: (a) the ratio image of  $TM3/TM1$ ; (b) the ratio image of  $TM4/TM3$ ; (c) the ratio image of  $TM5/TM7$ ; and (d) the ratio colour composite of  $TM5/TM7$  in blue,  $TM4/TM3$  in green and  $TM3/TM1$  in red

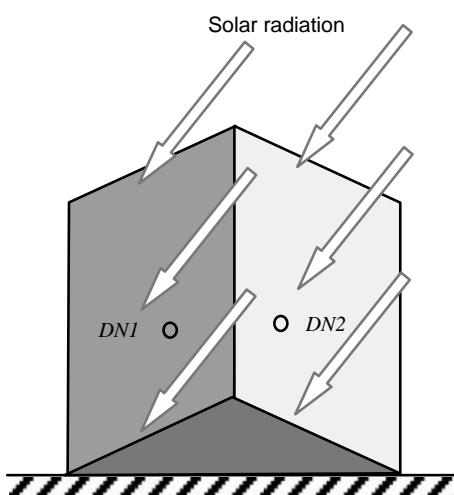
similarity between differencing and ratio techniques for selective enhancement. Many indices, such as the normalized difference vegetation index (NDVI), have been developed based on both differencing and ratio operations.

Ratio is also well known as an effective technique for suppressing topographic shadows. For a given incident angle of solar radiation, the radiation energy received by a land surface depends on the angle between the land surface and the incident radiation. Therefore, solar illumination on a land surface varies with terrain slope and aspect, which results in topographic shadows. In a remotely sensed image, the spectral information is often occluded by sharp variations of topographic shadowing. The DNs in different spectral bands of a multi-spectral image are proportional to the solar radiation received by the land surface and its spectral reflectance. Let  $DN(\lambda)$  represent the digital number of a pixel in an image of spectral band  $\lambda$ . Then

$$DN(\lambda) = \rho(\lambda)E(\lambda) \quad (3.9)$$

where  $\rho(\lambda)$  and  $E(\lambda)$  are the spectral reflectance and solar radiation of spectral band  $\lambda$  received at the land surface corresponding to the pixel.

As shown in Figure 3.5, suppose a pixel representing a land surface facing the Sun receives  $n$  times the radiation energy of that received by another



**Figure 3.5** Principle of shadow suppression function of ratio images

pixel of land surface facing away from the Sun; then the DNs of the two pixels in spectral bands  $i$  and  $j$  are as follows:

Pixel in shadow:  $DN1(i) = \rho(i)E(i)$  and  $DN1(j) = \rho(j)E(j)$

Pixel facing illumination:  $DN2(i) = n\rho(i)E(i)$  and  $DN2(j) = n\rho(j)E(j)$ .

Thus the ratio between band  $i$  and  $j$  for both pixels will be

$$\begin{aligned} R_{1_{i,j}} &= \frac{DN1(i)}{DN1(j)} = \frac{\rho(i)E(i)}{\rho(j)E(j)} \\ R_{2_{i,j}} &= \frac{DN2(i)}{DN2(j)} = \frac{n\rho(i)E(i)}{n\rho(j)E(j)} = \frac{\rho(i)E(i)}{\rho(j)E(j)}. \end{aligned} \quad (3.10)$$

Therefore,  $R_{1_{i,j}} = R_{2_{i,j}}$ .

Equations in (3.10) indicate that band ratios are independent of the variation of solar illumination caused by topographic shadowing and are decided only by the spectral reflectance of the image pixels. The pixels of the objects with the same spectral signature will result in the same band ratio values no matter whether they are under direct illumination or in shadow. Unfortunately, the real situation is more complicated than this simplified model because of atmospheric effects that often add different constants to different spectral bands. This is why the ratio technique can suppress topographic shadows but may not be able to eliminate their effects completely. Shadow suppression means losing topography that often accounts for more than 90% information of a multi-spectral image; ratio images therefore reduce SNRs significantly.

### 3.5 Index derivation and supervised enhancement

Infinite combinations of algebraic operations can be derived from basic arithmetic operations and algebraic functions. Aimless combinations of algebraic operations may mean an endless and potentially fruitless game; that is, you may spend a very long time without achieving any satisfactory result. Alternatively, you may happen upon a visually

impressive image without being able to explain or interpret it. To design a meaningful and effectively combined operation, knowledge of the spectral properties of targets is essential. The formulae should be composed on the basis of spectral or physical principles, and designed for the enhancement of particular targets; these are then referred to as *spectral indices*, such as the NDVI. An index can be considered as supervised enhancement. Here we briefly introduce a few commonly used examples of indices based on Landsat TM/ETM+ image data. You may design your own indices for a given image processing task based on spectral analysis. In Part Three of this book, you will find several examples of this kind of supervised enhancement in the teaching and research case studies.

### 3.5.1 Vegetation indices

As shown in Figure 3.6, healthy vegetation has a high reflection peak in the NIR and an absorption trough in the red. If we could see NIR, vegetation would be NIR rather than green. This significant difference between red and NIR bands is known as the *red edge*: it is a unique spectral property that makes vegetation different from all other ground objects. Obviously, this diagnostic spectral feature of vegetation can be very effectively enhanced by differencing and ratio operations. Nearly all the vegetation indices are designed to highlight the red edge in one way or another.

The NDVI is one of the most popular vegetation indices:

$$\text{NDVI} = \frac{\text{NIR} - \text{Red}}{\text{NIR} + \text{Red}}. \quad (3.11)$$

This index is essentially a difference between the NIR and red spectral band images. The summation of NIR and red in the denominator is a factor to normalize the NDVI to a value range  $[-1,1]$ .

The NDVI for TM imagery is

$$Y = \frac{\text{TM4} - \text{TM3}}{\text{TM4} + \text{TM3}}. \quad (3.12)$$

Vegetation can also be enhanced using a ratio index:

$$Y = \frac{\text{NIR} - \min(\text{NIR})}{\text{Red} - \min(\text{Red}) + 1} \quad (3.13)$$

and so for TM images

$$Y = \frac{\text{TM4} - \min(\text{TM4})}{\text{TM3} - \min(\text{TM3}) + 1}. \quad (3.14)$$

The effect of the subtraction of the band minimum is roughly to remove the added constants of atmospheric scattering effects (refer to the recommended remote sensing textbooks) so as to improve topography suppression by ratio. The value of 1 added to the denominator is to avoid a zero value.

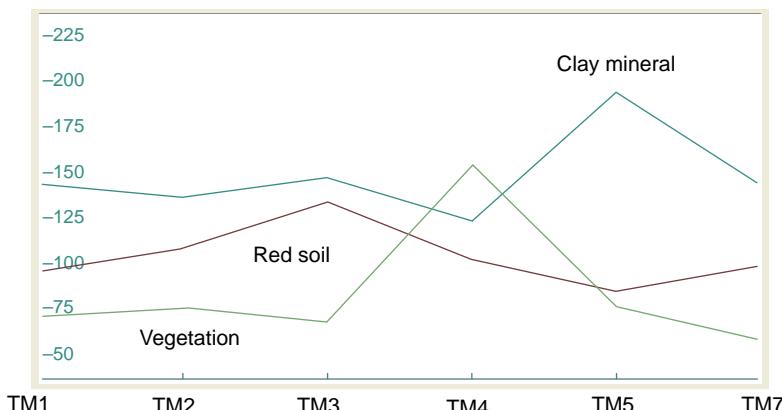
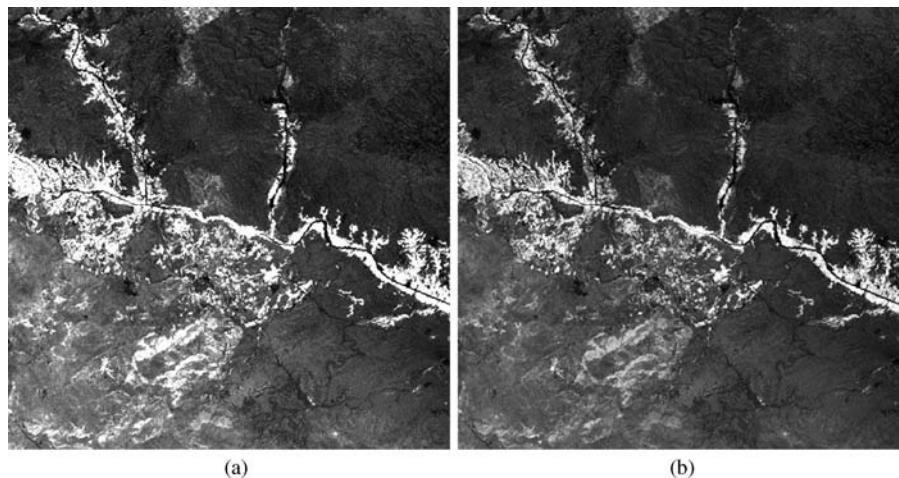


Figure 3.6 Image spectral signatures of vegetation, red soil and clay minerals



**Figure 3.7** (a) Landsat TM NDVI; and (b) vegetation ratio images

Figure 3.7 illustrates these vegetation indices derived from TM images.

### 3.5.2 Iron oxide ratio index

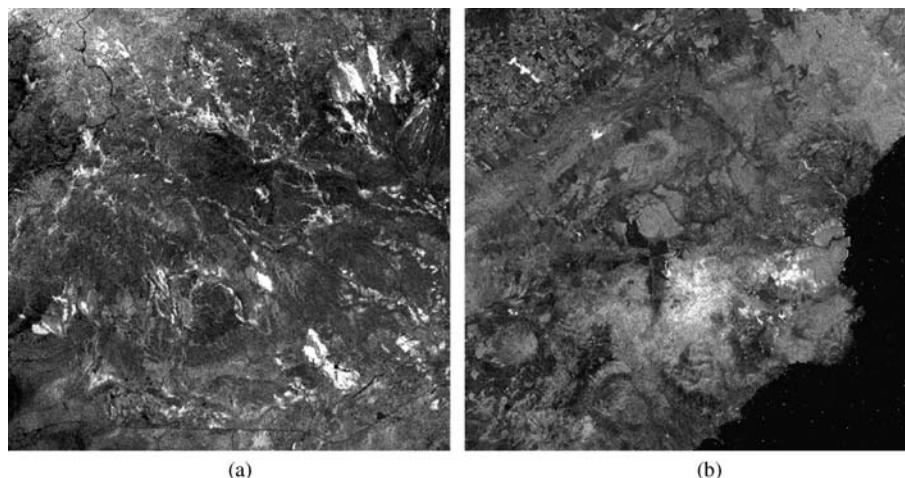
Iron oxides and hydroxides are some of the most commonly occurring minerals in the natural environment. They appear as red or reddish brown to the naked eye, because of high reflectance in the red and absorption in the blue (Figure 3.6). Typical red features on land surfaces, such as red soils, are

closely associated with the presence of iron-bearing minerals. We can enhance iron oxides using the ratio between the red and blue spectral band images (Figure 3.8a):

$$Y = \frac{\text{Red} - \min(\text{Red})}{\text{Blue} - \min(\text{Blue}) + 1}. \quad (3.15)$$

For TM imagery

$$Y = \frac{TM3 - \min(TM3)}{TM1 - \min(TM1) + 1}. \quad (3.16)$$



**Figure 3.8** Images of (a) TM iron oxide ratio index; and (b) TM clay mineral ratio index

### 3.5.3 TM clay (hydrated) mineral ratio index

Clay minerals are characteristic of hydrothermal alteration in rocks and are therefore very useful indicators for mineral exploration using remote sensing. The diagnostic spectral feature of clay minerals, which differentiates them from unaltered rocks, is that they all have strong absorption in the spectral range around  $2.2\text{ }\mu\text{m}$  (corresponding to TM band 7) in contrast to high reflectance in the spectral range around  $1.6\text{ }\mu\text{m}$  (corresponding to TM band 5) as shown in Figure 3.6. Thus clay minerals can be generally enhanced by the ratio between these two SWIR bands (Figure 3.8b):

$$Y = \frac{\text{TM}5 - \min(\text{TM}5)}{\text{TM}7 - \min(\text{TM}7) + 1}. \quad (3.17)$$

This index can only achieve a general enhancement of all clay minerals using TM or ETM+ images. ASTER imagery, on the other hand, offers five SWIR bands and enables more specific discrimination of different clay minerals (though still not specific identification). You might try to design ASTER indices to target various clay minerals by yourselves.

## 3.6 Standardization and logarithmic residual

A typical example of a combined algebraic operation is the so-called *standardization*:

$$Y_i = \frac{X_i}{\frac{1}{k} \sum_{\lambda=1}^k X_{\lambda}} \quad (3.18)$$

where  $X_i$  represent the band  $i$  image,  $Y_i$  the standardized band  $i$  image and  $k$  the total number of spectral bands.

This ratio-type operation can suppress topographic shadows based on the same principle as explained in Section 3.4. The denominator in the formula is the average image of all the bands of a multi-spectral image; this allows the ratio for every band to be produced using the same divisor. The standardization enables the spectral variation among different bands, at each pixel, to be better enhanced using the ratio to the same denominator.

If we consider (3.18) as an arithmetic-mean-based standardization then another technique called the *logarithmic residual* (Green and Craig, 1985) can be considered as a geometric-mean-based standardization:

$$\ln(R_{i\lambda}) = \ln(x_{i\lambda}) - \ln(x_{i\cdot}) - \ln(x_{\cdot\lambda}) + \ln(x_{\cdot\cdot}) \quad (3.19)$$

where  $x_{i\lambda}$  is the DN of pixel  $i$  in band  $\lambda$ ,  $x_{i\cdot} = (\prod_{\lambda=1}^k x_{i\lambda})^{1/k}$  is the geometric mean of pixel  $i$  over all the  $k$  bands,  $x_{\cdot\lambda} = (\prod_{i=1}^n x_{i\lambda})^{1/n}$  is the geometric mean of band  $\lambda$ , and  $x_{\cdot\cdot} = (\prod_{i=1}^n \prod_{\lambda=1}^k x_{i\lambda})^{1/kn}$  is the global geometric mean of all the pixels in all the bands.

Then,

$$y_{i\lambda} = e^{\ln(R_{i\lambda})} \quad (3.20)$$

where  $y_{i\lambda}$  is the logarithmic residual of  $x_{i\lambda}$ .

We can rewrite (3.19) in the following form:

$$\ln(R_{i\lambda}) = \ln \frac{x_{i\lambda}}{x_{i\cdot}} + \ln \frac{x_{\cdot\cdot}}{x_{\cdot\lambda}}. \quad (3.21)$$

The first term in (3.21) has a similar form to (3.18) but the denominator is a geometric mean instead of an arithmetic mean for pixel  $i$  over  $k$  bands. The second term in (3.21) is equivalent to a band spectral offset; it is constant for all the pixels in one spectral band but varies with different spectral bands.

The logarithmic residual technique suppresses topographic shadows more effectively than other techniques but the resulting images are not often visually impressive, even after a proper stretch, because of their rather low SNR.

## 3.7 Simulated reflectance

Many image processing techniques have been developed on the basis of fairly sophisticated physical or mathematical models but the actual constituent operations are simple arithmetic operations. Simulated reflectance technique (Liu *et al.*, 1997) is an example of this type.

### 3.7.1 Analysis of solar radiation balance and simulated irradiance

Suppose the solar radiation incident upon a solid land surface, of unit area equivalent to an image

pixel, is irradiance  $E$ . This energy is partially reflected and absorbed by the terrain material depending on the reflectance (or albedo)  $\rho$  and absorptance  $\alpha$ :

$$M_r = \rho E \quad (3.22)$$

$$M_a = \alpha E \quad (3.23)$$

where  $M_r$  is the reflected solar radiation and  $M_a$  the absorbed.

Considering the land surface as the surface of a solid medium of considerable thickness (which is generally true), and to satisfy the conservation of energy, we have

$$\rho + \alpha = 1. \quad (3.24)$$

Based on the concept of the radiation balance (Robinson, 1966) the solar radiation balance,  $B$ , on the Earth is described by

$$\begin{aligned} B &= E(1 - \rho) - M_e \\ &= \alpha E - M_e \\ &= M_a - M_e \end{aligned} \quad (3.25)$$

where  $M_e$  is the radiation emitted (thermal emission) from the land surface.

Then,

$$E = \rho E + \alpha E = M_r + M_a = M_r + M_e + B$$

or

$$E - B = M_r + M_e. \quad (3.26)$$

A dark (low-albedo) ground object absorbs more solar radiation energy (mostly in the visible to NIR spectral region) than a bright (high-albedo) object and eventually re-emits more thermal radiation in the thermal spectral region of 8–14  $\mu\text{m}$  because of the complicated thermodynamic processes within the terrain material. This general complementary relationship between reflected radiation  $M_r$  and emitted radiation  $M_e$  from the land surface can be easily observed in TM or ETM+ images in which dark subjects in visible band images are bright in thermal band images and vice versa. The phenomenon implies that the sum of  $M_r$  and  $M_e$ , the right side of Equation (3.26), can be treated roughly as a constant for a given irradiance  $E$ , and therefore is independent of the spectral properties (albedo) of the land surface.

Irradiance  $E$  varies with topography only. Suppose that the Sun is a ‘parallel’ radiation source to the Earth with constant incident radiant flux density  $M_s$ ; then the solar irradiance upon the land surface,  $E$ , varies with the angle between the land surface and the incident solar radiation,  $\gamma$ . When the land surface is perpendicular to the incident solar radiation  $M_s$ ,  $E$  is at its maximum and equal to  $M_s$ . If the solar radiation has an incident angle  $\theta_1$  (from nadir) and azimuth angle  $\phi_1$ , then the irradiance upon a land surface with slope angle  $\theta_2$  and aspect direction  $\phi_2$  can be calculated as

$$\begin{aligned} E &= M_s \sin \gamma \\ &= M_s [\sin \theta_1 \sin \theta_2 \cos(\phi_1 - \phi_2) + \cos \theta_1 \cos \theta_2]. \end{aligned} \quad (3.27)$$

As stated in (3.25),  $B$  is dependent on solar irradiance  $E$  and is therefore affected by topography in the same way as (3.27). Thus  $E - B$  varies mainly with topography while invariant to land surface albedo  $\rho$ . We refer to  $E - B$  defined by (3.26) as the *simulated irradiance* as it behaves like irradiance but with reduced energy by  $B$ .

### 3.7.2 Simulated spectral reflectance image

For the reflected spectral radiation of a particular wavelength  $\lambda$ , (3.22) becomes

$$M_r(\lambda) = \rho(\lambda)E(\lambda) \quad (3.28)$$

where  $\lambda$  is the spectral wavelength and  $\rho(\lambda)$  the spectral reflectance, and

$$E = \int_0^{\infty} E(\lambda)d\lambda. \quad (3.29)$$

The albedo  $\rho$ , as the total reflectance, is the integral of reflected spectral radiation, over the entire spectral range, divided by the irradiance:

$$\rho = \frac{\int_0^{\infty} \rho(\lambda)E(\lambda)d\lambda}{E}. \quad (3.30)$$

We define the simulated spectral reflectance of band  $\lambda$  as

$$\rho_{\text{sim}}(\lambda) = \frac{M_r(\lambda)}{M_r + M_e} = \rho(\lambda) \frac{E(\lambda)}{E - B}. \quad (3.31)$$

The right side of this equation comprises two components, the reflectance  $\rho(\lambda)$  and the ratio of the spectral irradiance of band  $\lambda$  to the simulated irradiance:  $E(\lambda)/(E - B)$ . This irradiance ratio is approximately constant for all pixels in the image band  $\lambda$  because both  $E(\lambda)$  and  $E - B$  vary with topography in a similar way to that defined by formula (3.27). An image defined by (3.31) is therefore directly proportional to the spectral reflectance image by a constant factor and, as a result, topographic features are suppressed.

Similarly, a *simulated thermal emittance* is defined as

$$\epsilon_{\text{sim}}(\lambda) = \frac{M_e(\lambda)}{M_r + M_e} = \frac{M_e(\lambda)}{E - B}. \quad (3.32)$$

Many airborne sensor systems have both multi-spectral bands and thermal bands with the same spatial resolution, such as airborne thematic mapper (ATM) images. In this case the simulated reflectance can be derived from these bands without degrading the spatial resolution. For ATM images,  $M_e$  is recorded in a broad thermal band  $ATM11$  (8–14  $\mu\text{m}$ ) and  $M_r$  is split and recorded into 10 reflective spectral bands  $ATM1$ – $ATM10$ . A simulated panchromatic band image  $M_r$  can therefore be generated as the weighted sum of the 10 reflective spectral bands:

$$M_r = \sum_{i=1}^{10} w_i ATM_i.$$

The weights  $w_i$  can be calculated either from the sensor gain factors and offsets, or using the solar radiation curve and the spectral bandwidths, as described later.

Thus, based on Equation (3.31), we have an ATM simulated reflectance image

$$\rho_{\text{sim}}(\lambda) = \frac{M_r(\lambda)}{E - B} = \frac{ATM\lambda}{ATM11 + \sum_{i=1}^{10} w_i ATM_i} \quad (3.33)$$

where  $\lambda = 1$ –10.

And a broadband thermal emittance image ( $\epsilon_{\text{sim}}$ ) for ATM is given by

$$\epsilon_{\text{sim}} = \frac{M_e}{E - B} = \frac{ATM11}{ATM11 + \sum_{i=1}^{10} w_i ATM_i}. \quad (3.34)$$

Similarly, we can derive the simulated reflectance/emittance images for Landsat TM, ETM+ and ASTER datasets but with degraded spatial resolution because the thermal band resolution of these sensor systems is significantly lower than that of reflective multi-spectral bands.

The TM simulated reflectance/emittance is

$$TM_{\text{sim},\rho,\epsilon}(\lambda) = \frac{TM\lambda}{TM6 + \sum_{i=1}^{1-5,7} w_i TM_i} \quad (3.35)$$

where  $TM_{\text{sim},\rho,\epsilon}$  is the simulated reflectance  $\rho_{\text{sim}}(\lambda)$  for  $\lambda = 1$ –5,7 and simulated emittance  $\epsilon_{\text{sim}}(\lambda)$  for  $\lambda = 6$ .

For ETM+ images we can use the same formula as above. We can also use the panchromatic band (ETM+ Pan) image to replace bands 2, 3 and 4 as the spectral range of the ETM+ Pan covers the same range of these three bands.

The ASTER simulated reflectance/emittance is

$$ASTER_{\text{sim},\rho,\epsilon}(\lambda)$$

$$= \frac{ASTER\lambda}{\sum_{i=1}^9 w_i ASTER_i + \sum_{j=10}^{14} w_j ASTER_j} \quad (3.36)$$

where  $ASTER_{\text{sim},\rho,\epsilon}$  is the simulated reflectance  $\rho_{\text{sim}}(\lambda)$  for  $\lambda = 1$ –9 and simulated emittance  $\epsilon_{\text{sim}}(\lambda)$  for  $\lambda = 10$ –14.

### 3.7.3 Calculation of weights

As described above, the simulated panchromatic ATM image is generated from a weighted sum of all the spectral bands of the ATM. In practice, this involves image pre-processing and calculation of weights and it can be carried out in various ways:

1. The standard decalibration procedure to convert the image DN in each spectral band to radiance using sensor gain and offset,

$$\text{Radiance} = \frac{\text{DN}}{\text{Gain}} - \text{Offset}.$$

The same conversion should also be performed on the thermal band. Careful atmospheric correction is needed before the summation is implemented.

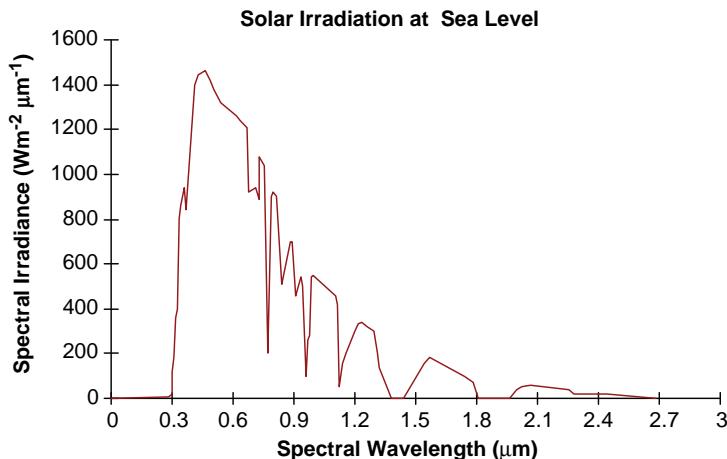


Figure 3.9 Solar spectral irradiation at sea level. After Fraster (1975)

- Use of the solar radiation curve (Figure 3.9) to calculate the weights. The average height of a spectral band in the solar radiation curve is measured and then multiplied by the bandwidth. The product, after rescaling to a percentage based on the summation of the products for all spectral bands, is the weight for the band. Each image band should be linearly stretched with a proper cut-off at both high and low ends of the histogram before the weight is applied. This stretch roughly removes the effects of atmospheric scattering and makes effective use of the whole DN range of the image (0–255 for 8 bit data). With all the image bands having the same DN range after the stretch, the weights calculated from the solar radiation curve can then be applied to all bands on an equal basis. The resultant simulated panchromatic image should, in principle, be further rescaled to optimize the albedo cancellation in the summation with the thermal band image so as to generate the simulated irradiance image (the denominator of Equation (3.34)). A real irradiance image should represent variation in topography only, without the influence of albedo. Effective cancellation of albedo in the simulated irradiance image is the key factor for retaining albedo information in the subsequent simulated reflectance image. In practice, the weights can be modified arbitrarily to achieve enhancement of particular spectral features. Tables 3.1 and 3.2 gives the weights of the ATM and ETM+ bands.

### 3.7.4 Example: ATM simulated reflectance colour composite

Figure 3.10 shows a colour composite of ATM bands 9, 4 and 2 in red, green and blue (a) and the simulated reflectance colour composite of the same bands (b). The normal colour composite was prepared using the BCET for optimum colour presentation. For the simulated reflectance colour composite, all image bands were linearly stretched with clipping at the both ends of the image histograms to make full use of the 8 bit (0–255) value range and then the weights in Table 3.1 were applied. The resultant simulated reflectance images were linearly rescaled to 0–255 for colour display.

The colour composite of ATM bands 9, 4 and 2 is generally good for visual interpretation of rock types, alteration minerals and soil/regolith. Figure 3.10 gives the overall impression that the simulated reflectance colour composite (b) has a very similar colour appearance to the normal colour composite (a) but with topography suppressed. This general similarity makes the visual interpretation easy since the image colours relate directly to spectral signatures. Further examination indicates that the simulated reflectance image has more spectral (colour) variations than the normal colour composite. In the simulated reflectance colour composite, the main contribution to image contrast is given by spectral variation rather than topography. This enables the spectral features within very light or

**Table 3.1** Weights for ATM bands in calculation of simulated reflectance

ATM band	1	2	3	4	5	6	7	8	9	10	11
Wavelength (μm)	0.42–0.45	0.45–0.52	0.52–0.60	0.60–0.62	0.63–0.69	0.69–0.75	0.76–0.90	0.91–1.05	1.55–1.75	2.08–2.35	8.50–14.0
E (W/m <sup>2</sup> /μm)	1440	1420	1307	1266	1210	909	710	505	148	43	
Weight	0.07	0.16	0.17	0.04	0.12	0.09	0.16	0.12	0.05	0.02	1

**Table 3.2** Weights for TM/ETM+ bands in calculation of simulated reflectance

TM/ETM+ band	1	2	3	4	5	6	7
Wavelength (μm)	0.45–0.53	0.52–0.60	0.63–0.69	0.76–0.90	1.55–1.75	10.4–12.5	2.08–2.35
Weight	0.2	0.3	0.2	0.1	0.1	1	0.1

very dark (low-albedo or topographic shadow) areas to be enhanced effectively.

### 3.7.5 Comparison with ratio and logarithmic residual techniques

As mentioned before, many techniques, such as ratio, standardization and logarithmic residual, have been developed to enhance information relating to the spectral reflectance of ground objects by suppressing topographic shadowing. All these techniques suffer the limitation that albedo variation is suppressed together with topographic shadow. These techniques cannot, therefore, effectively differentiate objects with similar spectral profiles but different albedos, such as separation between black and grey. This is because they involve ratio-type operations either between the spectral reflectance of different bands or between band spectral reflectance and albedo.

$$\text{Ratio : } R_{ij} = \frac{\text{Band}(i)}{\text{Band}(j)} = \frac{\rho(i)E(i)}{\rho(j)E(j)}.$$

The spectral irradiance ratio  $E(i)/E(j)$  is constant for any pixel in the image so that topographic shading is removed. The reflectance ratio  $\rho(i)/\rho(j)$  varies with the spectral signatures of the pixels but cancels out albedo variation because a high-albedo object will have a similar reflectance ratio to that of a low-albedo object if the two have similar spectral profiles. In other words, the band ratio technique cannot separate albedo from irradiance on the land surface and, consequently, the method suppresses the variation of both.

*Logarithmic Residual:* Formula (3.19) of logarithmic residual operation can be rewritten as

$$R_{i\lambda} = \frac{x_{i\lambda}/x_i}{x_{.\lambda}/x_{..}} = \frac{\rho_i(\lambda)E_i(\lambda)/\rho_i E_i}{\rho_{..}(\lambda)E(\lambda)/\rho_{..} E} \quad (3.37)$$

where,

$\rho_i(\lambda)$  is the spectral reflectance of band  $\lambda$  of pixel  $i$   
 $\rho_i$  is the albedo of pixel  $i$   
 $\rho_{..}(\lambda)$  is the average reflectance of the spectral band  $\lambda$   
 $\rho_{..}$  is the average albedo of the whole image  
and the other variables are as defined in (3.19).



**Figure 3.10** (a) The colour composite of ATM bands 9, 4 and 2 in red, green and blue; and (b) the simulated reflectance colour composite of ATM bands 9, 4 and 2

The irradiance ratio  $E_i(\lambda)/E_i$  is constant for all pixels and thus independent of position  $i$ ,

$$\frac{E_i(\lambda)}{E_i} = \frac{E(\lambda)}{E}.$$

Thus Equation (3.37) can be simplified as

$$R_{i\lambda} = \frac{\rho_{..}}{\rho_{.}(\lambda)} \times \frac{\rho_i(\lambda)}{\rho_i}. \quad (3.38)$$

Formula (3.38) is irrelevant to irradiance and so topographic shadows are therefore eliminated by logarithmic residual processing. This formula is actually a product of two ratios: the ratio of the average albedo of the whole image against the average reflectance of spectral band  $\lambda$ ,  $\rho_{..}/\rho_{.}(\lambda)$ ; and the ratio of pixel spectral reflectance and pixel albedo,  $\rho_i(\lambda)/\rho_i$ . The ratio  $\rho_{..}/\rho_{.}(\lambda)$  is independent of position and constant for all the pixels in the band  $\lambda$  logarithmic residual image. The variation in a logarithmic residual image is therefore controlled only by the ratio  $\rho_i(\lambda)/\rho_i$ . In a similar way to ratio images, this ratio cancels the variation of albedo.

The advantage of the simulated reflectance technique is that it does not involve ratio operations between spectral reflectances or between spectral reflectances and albedo. By using the thermal band image to generate a simulated irradiance component, the simulated reflectance technique suppresses topographic shadows but retains albedo information. Thus the spectral information can be better enhanced.

### 3.8 Summary

In this chapter, we learned about simple arithmetic operations between images and discussed their main applications in image spectral enhancement. The key point to understand is that all image algebraic operations are point based and performed among the corresponding pixels in different images without the involvement of neighbouring pixels. We can therefore regard algebraic operations as *multi-image point operations*.

A major application of image algebraic operations is the selective enhancement of the spectral signatures

of intended targets in a multi-spectral image. For this purpose, investigating the spectral properties of these targets is essential to the composition of effective algebraic operations; random attempts are unlikely to be fruitful. This procedure, from spectral analysis to composing an algebraic formula, is generally referred to as *supervised enhancement*. If such a formula is not image dependent and so can be widely used, it is called an *index* image; for instance, the NDVI is a very well-known vegetation index image.

One important issue for spectral enhancement is the suppression of topographic shadowing effects. The ratio-based techniques, standardization and logarithmic residual, are based on numerical cancellation of variations in solar illumination. The results are related to ratios of spectral reflectance between different bands while variation in albedo will be removed together with topography. The simulated reflectance, in contrast, which is based on a simplified physical model of solar radiation on the land surface, represents all the properties of true spectral reflectance, including albedo, with only a constant difference.

## Questions

- 3.1 Why is an image algebraic operation also known as a multi-image point operation? Write down the mathematical definition of the multi-image point operation.
- 3.2 Why can image addition improve the image SNR? If nine photographs of the same scene are taken, using a stationary camera under identical illumination conditions, and then summed to generate an average image, by how many times is the SNR is improved in comparison with any individual picture?
- 3.3 Describe image difference (subtraction) and ratio (division) operations and then compare the two techniques in terms of change detection, selective enhancement and processing efficiency.

- 3.4 How important are the weights in image subtraction? Suggest a suitable pre-processing technique for image differencing.
- 3.5 Why does image differencing decrease the SNR?
- 3.6 Describe image multiplication and its main application.
- 3.7 Explain the characteristics of the value range of a ratio image. Do you think that two reciprocal ratio images contain the same information when displayed after linear scale and, if so, why?
- 3.8 Using a diagram, describe a ratio image in terms of a coordinate transformation between Cartesian and polar coordinates.
- 3.9 Explain the principles of topographic suppression using the image ratio technique.
- 3.10 What is NDVI and how is it designed?
- 3.11 Describe the design and functionality of the Landsat TM or ETM + iron oxide and clay indices.
- 3.12 Try the normalized differencing approach, similar to NDVI, to enhance iron oxide and clay minerals. Compare the results with the relevant ratio indices and explain why the ratio-based indices are more effective for these two minerals. (Key: The red edge signal for vegetation is much stronger than the difference between red and blue for iron oxide between the SWIR bands for clay minerals.)
- 3.13 Describe and compare the standardization and logarithmic residual techniques and their functionalities.
- 3.14 What is simulated reflectance? What is the essential condition for the derivation of a simulated reflectance image?
- 3.15 Referring to the physical model for the derivation of simulated reflectance, explain why it can be regarded as a true simulation of reflectance.
- 3.16 What are the major advantages of simulated reflectance over the ratio, standardization and logarithmic residual techniques?



# 4

## Filtering and Neighbourhood Processing

Filtering is a very important research field of digital image processing. All filtering algorithms involve so-called *neighbourhood processing* because they are based on the relationship between neighbouring pixels rather than a single pixel in point operations.

Digital filtering is useful for enhancing lineaments that may represent significant geological structures such as faults, veins or dykes. It can also enhance image texture for discrimination of lithologies and drainage patterns. For land-use studies, filtering may highlight the textures of urbanization, road systems and agricultural areas and, for general visualization, filtering is widely used to sharpen images. However, care should be taken because filtering is not so ‘honest’ in retaining the information of the original image. It is advisable to use filtered images in close reference to colour composite images or black and white single band images for interpretation.

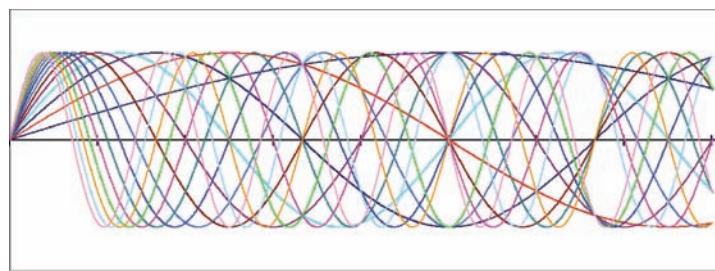
Digital filtering can be implemented either by ‘box filters’ based on the concept of convolution in the spatial domain or using the Fourier transform (FT) in the frequency domain. In the practical applications of remote sensing, convolution-based box filters are the most useful for their computing efficiency and reliable results. In giving a clear explanation of the physical and mathematical meanings of filtering, the FT is essential for understanding the principle of convolution. The FT is less computationally efficient for raster data, in terms of speed and computing resources, but it is

more versatile than convolution in accommodating various filtering functions.

For point operations, we generally regard an image as a raster data stream and denote  $x_{ij} \in X$  as a pixel at line  $i$  and column  $j$  in image  $X$ . As the pixel coordinates in an image are irrelevant for point operations, the subscripts  $ij$  are not involved in the processing and can be ignored. For filtering and neighbourhood processing, however, the pixel coordinates are very relevant and, in this sense, we regard an image as a 2D function. We therefore follow the convention of denoting an image with a pixel at image column  $x$  and line  $y$  as a 2D function  $f(x, y)$ , when introducing essential mathematical concepts of filtering in this chapter. On the other hand, for simplicity, the expression of  $x_{ij} \in X$  is still used in describing some filters and algorithms.

### 4.1 Fourier transform: understanding filtering in image frequency

The information in an image can be considered as the spatial variations at various frequencies or the assembly of spatial information of various frequencies as illustrated in Figure 4.1. Smooth, gradual tonal variations represent low-frequency information while sharp boundaries represent high-frequency information. The FT is a powerful tool for converting image information from the spatial domain into the



**Figure 4.1** An image can be considered as an assembly of spatial information at various frequencies

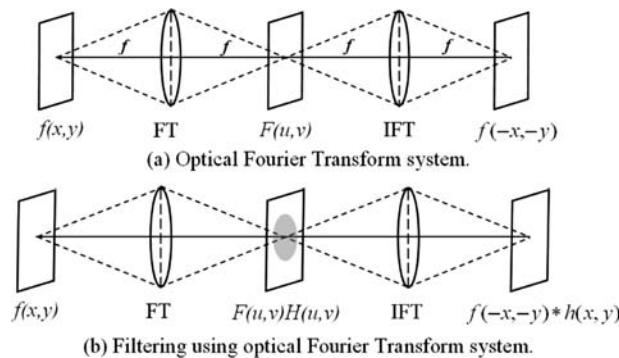
frequency domain. Filtering can thus be performed on selected frequencies and this is particularly useful for removing periodical noise, such as that induced by scanning lines.

Firstly, let us try to understand the physical meaning of the FT based on a simple optical filtering system before we start the slightly boring mathematics. As shown in Figure 4.2, it is a so-called 4f optical system. The  $f$  is the focal length of the lens performing the FT and inverse Fourier transform (IFT). Given an image  $f(x, y)$ , where  $x$  and  $y$  are the spatial coordinates of pixel position, the first lens performs an FT to transform  $f(x, y)$ , at its front focal plane, to a Fourier transformation  $F(u, v)$ , a frequency spectrum with frequencies  $u$  in the horizontal and  $v$  in the vertical direction, at its rear focal plane. The second lens then performs an IFT to transform  $F(u, v)$  at its front focal plane back to the image at its rear focal plane but with a 180° rotation  $f(-x, -y)$ . Figure 4.3 shows an image (a) and its FT spectrum plane (b); the frequency increases from zero at the centre of the spectrum plane to higher and higher frequencies towards the edge of the plane. If a filter (mask)  $H(u, v)$  is placed

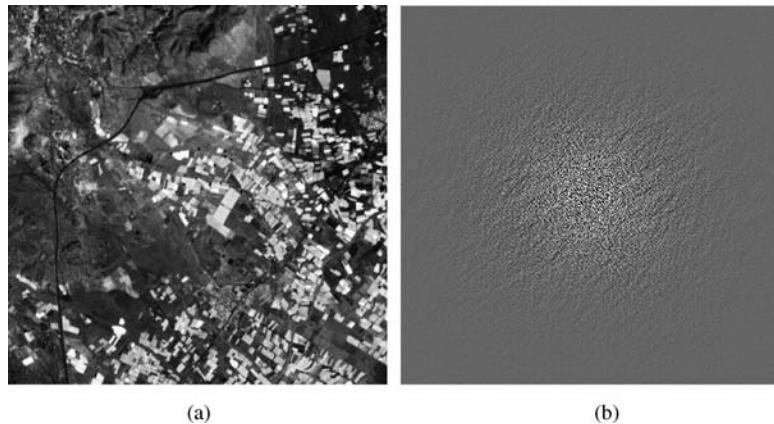
at the rear focal plane of the FT lens, to mask off the signals of particular frequencies in  $F(u, v)$ , which is equivalent to the operation  $F(u, v)H(u, v)$ , then a filtered image  $g = f(-x, -y)^*h(x, y)$  (with 180° rotation) is produced, as shown in Figure 4.2b. Such an optical FT system can perform filtering very efficiently and with great flexibility since the filter can be designed in various arbitrary shapes which are very difficult or impossible to define by mathematical functions. Unfortunately, such an analogue approach is limited by the requirements of delicate and expensive optical instruments and very strict laboratory conditions. On the other hand, with rapid progress in computing technology and the development of the more efficient and accurate fast Fourier transform (FFT) algorithms, frequency domain filtering has become a common function for most image processing software packages.

From the illustration in Figure 4.2, we know that the FT-based filtering has three steps:

1. FT to transfer an image into the frequency domain.
2. Remove or alter the data of particular frequencies using a filter.



**Figure 4.2** Optical Fourier transform system for filtering



**Figure 4.3** (a) An image  $f(x, y)$  and (b) its FT frequency spectrum  $F(u, v)$

3. IFT to transfer the filtered frequency spectrum back to the spatial domain to produce a filtered image.

Suppose  $f(x, y)$  is an input image and  $F(u, v)$  is the 2D FT of  $f(x, y)$ ; then the continuous (C) forms of the 2D FT and IFT are

$$\text{C2DFT } F(u, v) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(x, y) e^{-i2\pi(ux+vy)} dx dy \quad (4.1)$$

$$\text{C2DIFT } f(x, y) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} F(u, v) e^{i2\pi(ux+vy)} du dv \quad (4.2)$$

where  $i = \sqrt{-1}$ .

The discrete (D) forms are

$$\text{D2DFT } F(u, v) = \frac{1}{n} \sum_{x=0}^{n-1} \sum_{y=0}^{n-1} f(x, y) e^{-i2\pi(ux+vy)/n} \quad (4.3)$$

$$\text{D2DIFT } f(x, y) = \frac{1}{n} \sum_{u=0}^{n-1} \sum_{v=0}^{n-1} F(u, v) e^{i2\pi(ux+vy)/n} \quad (4.4)$$

where  $n$  is the calculation window size.

The operations of the FT and IFT are essentially the same but one is from the image domain to the frequency domain and the other from the frequency domain to the image domain. An important property of the FT is known as the *convolution theorem*.

It states that if  $F$  and  $H$  are the Fourier transforms of function  $f$  and  $h$ , then the Fourier transform of the convolution  $f * h$  is equal to the product of  $F$  and  $H$ :

$$\text{FT}(f * h) = \text{FT}(f)\text{FT}(h) = FH. \quad (4.5)$$

The inverse form is

$$f * h = \text{IFT}[\text{FT}(f)\text{FT}(h)] = \text{IFT}(FH) \quad (4.6)$$

or

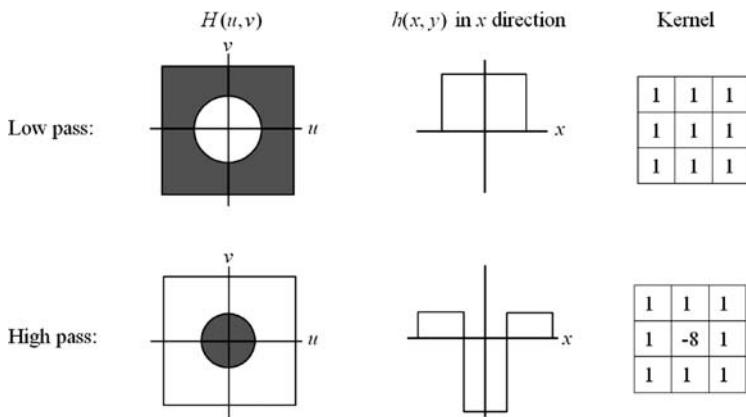
$$G = FH \text{ for } g = f * h \text{ and } G = \text{FT}(g). \quad (4.7)$$

This is the key concept of filtering based on the FT.  $F$  is the frequency representation of the spatial information of image  $f$ . If  $H$  is a filtering function to change the power of particular frequencies or make them zero, the formula (4.6) performs the filtering that changes or removes these frequencies and produces a filtered image  $g = f * h$ .

## 4.2 Concepts of convolution for image filtering

From the convolution theorem, we know that the image filtering using the FT is equivalent to a convolution between an image  $f(x, y)$  and a function  $h(x, y)$  that is usually called the *point spread function* (PSF). A 2D convolution is defined as

$$\begin{aligned} g(x, y) &= f(x, y) * h(x, y) \\ &= \int \int f(u, v) h(x-u, y-v) du dv. \end{aligned} \quad (4.8)$$



**Figure 4.4** Illustration of low-pass and high-pass frequency filters and their PSFs

Comparing (4.8) with the convolution theorem (4.6), it is clear that filtering in the image domain by a PSF defined as  $h(x, y)$  is equivalent to that in the frequency domain by a frequency filtering function  $H(u, v)$ . The  $h(x, y)$  is actually the FT or the ‘image’ of the frequency filtering function  $H(u, v)$ . Filtering can therefore be performed directly in the image domain by convolution without involving time-consuming FTs and IFTs, if the image presentation of a frequency filter can be found. For many standard frequency filtering functions, such as high-pass and low-pass filters, their images can be derived easily by the IFT, as illustrated in Figure 4.4. It is clear then that convolution is a shortcut for filtering operations.

In the case of discrete integer digital images, the integral form of (4.8) becomes a summation:

$$g(x, y) = \sum_{u=-\infty}^{+\infty} \sum_{v=-\infty}^{+\infty} f(u, v)h(x-u, y-v). \quad (4.9)$$

If the range over which the PSF  $h(x, y)$  is non-zero is  $(-w, +w)$  in one dimension and  $(-t, +t)$  in the other, then Equation (4.9) can be written as

$$g(x, y) = \sum_{u=x-w}^{x+w} \sum_{v=y-t}^{y+t} f(u, v)h(x-u, y-v). \quad (4.10)$$

In digital image filtering,  $w$  and  $t$  are the half size of a filter kernel in the horizontal and vertical directions. The pixel of a filtered image,  $g(x, y)$ , is created by a summation over the neighbourhood pixels  $f(u, v)$  surrounding the input image pixel  $f(x, y)$  weighted by  $h(x-u, y-v)$ .

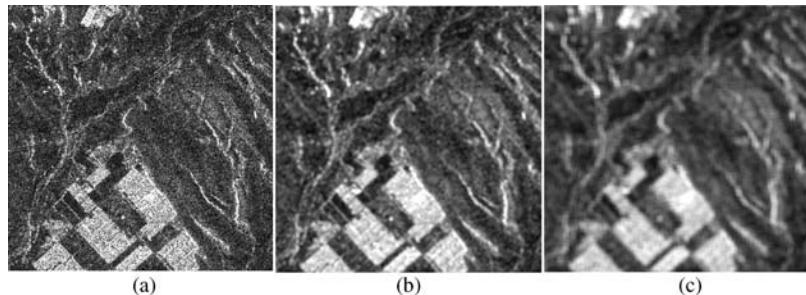
As explained before, the filter kernel, or the PSF  $h(x, y)$ , is a matrix or the image of a frequency filter  $H(u, v)$  according to the convolution theorem (4.6). The numbers in a filter kernel  $h(x, y)$  are the weights for summation over the neighbourhood of  $f(x, y)$ . For the whole image, the convolution filtering is performed by shifting the filter kernel, pixel by pixel, to apply (4.10) to every pixel of the image being filtered. Although the kernel size may be either an odd number or even number, an odd number is preferred to ensure the symmetry of the filtering process. A kernel of even-number size results in a half pixel shift in the filtering result. Commonly used filter kernel sizes are  $3 \times 3$ ,  $5 \times 5$  or  $7 \times 7$ . Rectangular kernels are also used according to particular needs.

Convolution is the theoretical foundation of image domain or spatial filtering, though many spatial filters are not necessarily based on the mathematical definition of convolution but on neighbourhood relationships.

For neighbourhood processing, in either domain, the image margins of half the size of the processing window can either be excluded from the output or be processed together with a mirror copy of the half window size block on the inside of the margins.

### 4.3 Low-pass filters (smoothing)

Smoothing filters are designed to remove high-frequency information and retain low-frequency



**Figure 4.5** (a) Original image; (b)  $5 \times 5$  mean filter result; and (c)  $9 \times 9$  mean filter result

information, thus reducing the noise but at the cost of degrading image detail. Figure 4.4 illustrates a typical low-pass filter  $H(u, v)$  and the corresponding PSF  $h(x, y)$ . Most kernel filters for smoothing involve weighted averages among the pixels within the kernel. The larger the kernel, the lower the frequency of information retained. Smoothing based on averaging is effective for eliminating noise pixels, which are often distinguished by very different DNs from their neighbours, but, on the other hand, the process blurs the image by also removing the high-frequency information. As illustrated in Figure 4.5, an SAR multi-look image appears noisy because of radar speckle (Figure 4.5a); the speckle is effectively removed using a  $5 \times 5$  mean filter, so better revealing the ground features (Figure 4.5b); these features are blurred when a  $9 \times 9$  mean filter is applied (Figure 4.5c). For removing noise without blurring images, *edge-preserved smoothing* becomes an important research topic of filtering.

The following are examples of  $3 \times 3$  low-pass filter kernels. The size and shape of the kernels can be varied.

*Mean filters:*

$$\frac{1}{9} \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix} \quad \frac{1}{5} \begin{pmatrix} 0 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 0 \end{pmatrix}.$$

*Weighted mean filters:*

$$\frac{1}{16} \begin{pmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{pmatrix} \quad \frac{1}{6} \begin{pmatrix} 0 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 0 \end{pmatrix}.$$

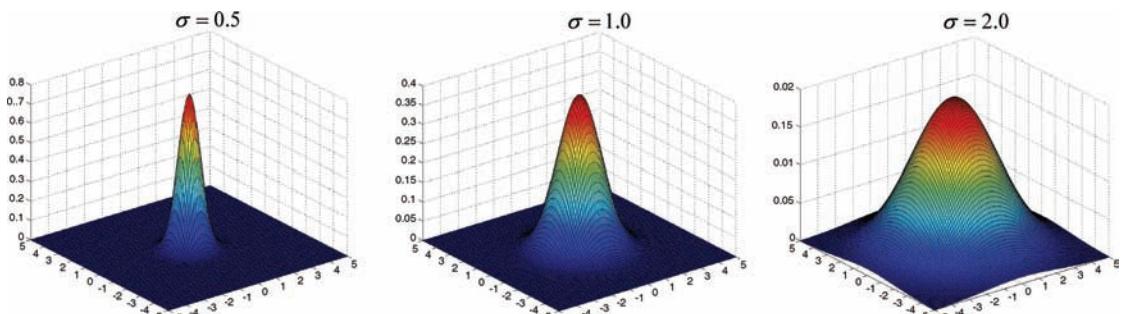
#### 4.3.1 Gaussian filter

A Gaussian filter is a smoothing filter with a 2D Gaussian function as its PSF:

$$G(x, y) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{x^2 + y^2}{2\sigma^2}\right) \quad (4.11)$$

where  $\sigma$  is the standard deviation of the Gaussian function.

Figure 4.6 presents the PSF of Gaussian filters with  $\sigma = 0.5, 1.0$  and  $2.0$ . The Gaussian function is a



**Figure 4.6** Three-dimensional representations of Gaussian filters with different  $\sigma$  values

continuous function. For discrete raster images, the Gaussian filter kernel is a discrete approximation for a given  $\sigma$ . For  $\sigma=0.5$ , a Gaussian filter is approximated by a  $3 \times 3$  kernel as

$$G_{\sigma=0.5} = \begin{pmatrix} 0.0113 & 0.0838 & 0.0113 \\ 0.0838 & 0.6193 & 0.0838 \\ 0.0113 & 0.0838 & 0.0113 \end{pmatrix}.$$

Obviously, it is essentially a weighted mean filter.

### 4.3.2 The $k$ nearest mean filter

This involves the reassignment of the value of a pixel  $x_{ij}$  of image  $X$  to the average of the  $k$  neighbouring pixels in the kernel window whose DNs are closest to that of  $x_{ij}$ . A typical value of  $k$  is 5 for a  $3 \times 3$  square window. This approach avoids extreme DNs, which are likely to be noise, and ensures their removal. On the other hand, if the pixel in the kernel window is an edge pixel, taking the average of  $k$  nearest DNs is more like to preserve the edge. The  $k$  nearest mean is therefore an edge-preserving smoothing filter. As shown in Figure 4.7a, the central pixel 0 is very likely to represent noise. The five DNs nearest to 0 are 0, 54, 55, 57 and 58 and the mean is 44.8. The suspected noise DN, 0, at the central pixel is then replaced with the  $k$  nearest mean 44.8 that is nearer to 0 than the average 53.4 produced by a mean filter. For the case in Figure 4.7b, the central pixel DN 156 is replaced with 158.6, the  $k$  nearest mean of 156, 155, 159, 161 and 162. As 158.6 is much nearer to 156 than the average 113 produced by a mean filter, the edge between the pixels in the DN range of 54–58 and those in 155–162 is better preserved and the image is less blurred.

55	58	65
57	0	63
54	61	68

(a)

55	58	155
57	156	159
54	161	162

(b)

**Figure 4.7** Two image templates in a  $3 \times 3$  kernel window to illustrate the effects of edge-preserving filters: (a) a template with a noise pixel 0; and (b) a template with an edge between DNs of 54–58 and 155–162

### 4.3.3 Median filter

Here the value of a pixel  $x_{ij}$  of image  $X$  is reassigned to the median DN of its neighbouring pixels in a kernel window (e.g.  $3 \times 3$ ). We use the same examples in Figure 4.7 to explain.

For image template (a), the DNs are ranked: 0, 54, 55, 57, **58**, 61, 63, 65, 68; the median in this neighbourhood is 58. The central pixel 0, a suspected isolated noise pixel, is replaced by the median 58 which is a reasonable estimate based on the neighbouring pixels.

For image template (b), the DNs are ranked: 54, 55, 57, 58, **155**, 156, 159, 161, 162; the central DN 156 is replaced with the median value 155 which is very close in value to the original DN. Thus the sharp edge between pixels with DNs in range of 54–58 and those with DNs ranging from 155 to 162 is preserved. If a mean filter were used, the central pixel DN would be replaced by a value of 113 which is significantly lower than the original DN 156 and, as a result, the edge would be blurred. The median filter is therefore an edge-preserving smoothing filter.

In summary, the median filter can remove isolated noise without blurring the image too much. Median filtering can be performed in the vertical or horizontal direction, if the filter window is one line or one column width instead of a square box.

### 4.3.4 Adaptive median filter

The adaptive median filter is designed from the basic principle of the median filter as follows:

Median1	Median2	Median3
x	x	x
X	x	X
x	x	x

It involves the reassignment of the value of a pixel  $x_{ij}$  of image  $X$  to the median of the above three medians in its  $3 \times 3$  neighbourhood. This filter is unlikely to change the DN of a pixel if it is not noise and thus is very effective for edge preservation.

Using the same examples in Figure 4.7, for image template (a), we have

Median1	Median2	Median3
55	65	58
0	57	0

Median1	Median2	Median3
54	68	61
		0

Thus, Median1 = 55, Median2 = 58 and Median3 = 0, and the median of these three medians is 55; thus the central DN 0 is replaced with 55 and the isolated noise is removed.

For image template (b), we have

Median1	Median2	Median3
55	155	58
156	57	156
54	162	159
		156

Here, Median1 = 155, Median2 = 156 and Median3 = 156, and the median of these three medians is 156; thus the central DN 156 remains unchanged and the edge is preserved. It is clear this is a strong edge-preserving smoothing filter. Larger window sizes can also be used with the adaptive median filter.

### 4.3.5 The $k$ nearest median filter

The design of this filter combines the principles of the  $k$  nearest mean filter and the median filter. It involves the reassignment of the value of a pixel  $x_{ij}$  of image  $X$  to the median of the  $k$  neighbour pixels in the kernel window whose DNs are closest to that of  $x_{ij}$ . A typical value of  $k$  is 5 for a  $3 \times 3$  square window. Taking the same example in Figure 4.7, for template (a), the five nearest DNs to the central pixel 0 are: 0, 54, 55, 57 and 58. The suspected noise DN, 0, at the central pixel is then replaced with the  $k$  nearest median 55. This is a more reasonable replacement value and is closer to the neighbourhood of  $x_{ij}$  than 44.8 generated by the  $k$  nearest mean filter. For template (b), the five nearest DNs to the central pixel 156 are: 155, 156, 159, 161 and 162.

Thus the central pixel DN 156 is replaced by 159. The  $k$  nearest median filter is more effective for removing noise and preserving edges than the  $k$  nearest mean filter.

### 4.3.6 Mode (majority) filter

This is a rather ‘democratic’ filter. A pixel value is reassigned to the most popular DN among its neighbourhood pixels. This filter performs smoothing based on the counting of pixels in the kernel rather than numerical calculations. Thus it is suitable for smoothing images of non-sequential data (symbols) such as classification images or other discrete raster data. For a  $3 \times 3$  kernel, the recommended majority number is 5. If there is no majority found within a kernel window, then the central pixel in the window remains unchanged.

For example:

6	6	6
6	2	6
5	5	6

There are six pixels here with DN = 6, therefore the central DN 2 is replaced by 6. For a classification image, the numbers in this window are the class numbers and their meaning is no different to class symbols A, B and C. If we use a mean filter, the average of the DNs in the window will be 5.3, but class 5.3 has no meaning in a classification image!

### 4.3.7 Conditional smoothing filters

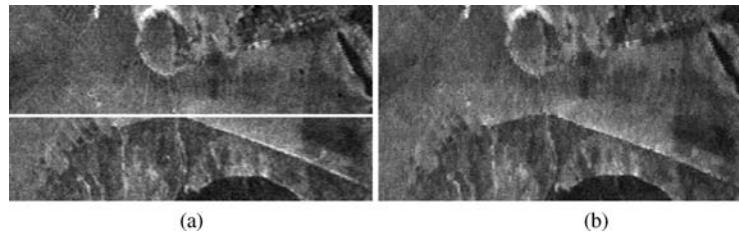
Filters of this type have the following general form:

---

```
If (some condition)
    Apply filter 1
Else
    Apply filter 2
Endif
```

---

Typical examples of this type of filters are *noise cleaning filters*. These filters are all designed based



**Figure 4.8** (a) An airborne thematic mapper (ATM) image is contaminated by several bad lines as the result of occasional sensor failure; and (b) the bad lines are successfully removed by the clean lines filter

on the assumption that a bad pixel, a bad line or a bad column (data missing or CCD malfunction) in an image will have significantly different DNs from its neighbourhood pixels, lines or columns.

Consider a  $3 \times 3$  window neighbourhood of pixel  $x_{ij}$ :

$$\begin{matrix} x_{i-1,j-1} & x_{i-1,j} & x_{i-1,j+1} \\ x_{i,j-1} & x_{i,j} & x_{i,j+1} \\ x_{i+1,j-1} & x_{i+1,j} & x_{i+1,j+1}. \end{matrix}$$

#### Clean pixels filter

The difference between a bad pixel and the mean of either of its two alternative neighbourhood pixels will be greater than the difference between the two neighbourhood means and thus can be identified and replaced by the mean of its four nearest pixels:

$$AVE1 = \frac{1}{4}(x_{i-1,j-1} + x_{i+1,j+1} + x_{i+1,j-1} + x_{i-1,j+1})$$

$$AVE2 = \frac{1}{4}(x_{i-1,j} + x_{i+1,j} + x_{i,j-1} + x_{i,j+1})$$

$$DIF = |AVE1 - AVE2|.$$

If :  $|x_{i,j} - AVE1| > DIF$  and  $|x_{i,j} - AVE2| > DIF$

then :  $y_{i,j} = AVE2$

otherwise:  $y_{i,j} = x_{i,j}$ .

#### Clean lines filter

If an image has a bad line, the difference between a pixel in this line and the mean of either the line above or below within the processing window will be greater than the difference between the means of these two lines. Then the bad line pixel can be replaced by the average of the pixels above and below it. As such, a bad line is replaced by the

average of the image line above and below it. Figure 4.8 illustrates the effects of the filter:

$$AVE1 = \frac{1}{3}(x_{i-1,j-1} + x_{i-1,j} + x_{i-1,j+1})$$

$$AVE2 = \frac{1}{3}(x_{i+1,j-1} + x_{i+1,j} + x_{i+1,j+1})$$

$$DIF = |AVE1 - AVE2|$$

If :  $|x_{i,j} - AVE1| > DIF$  and  $|x_{i,j} - AVE2| > DIF$

then :  $y_{i,j} = (x_{i-1,j} + x_{i+1,j})/2$

otherwise:  $y_{i,j} = x_{i,j}$ .

#### Clean columns filter

This filter is designed with the same logic as the above but in the vertical direction:

$$AVE1 = \frac{1}{3}(x_{i-1,j-1} + x_{i,j-1} + x_{i+1,j-1})$$

$$AVE2 = \frac{1}{3}(x_{i-1,j+1} + x_{i,j+1} + x_{i+1,j+1})$$

$$DIF = |AVE1 - AVE2|$$

If:  $|x_{i,j} - AVE1| > DIF$  and  $|x_{i,j} - AVE2| > DIF$

then :  $y_{i,j} = (x_{i,j-1} + x_{i,j+1})/2$

otherwise:  $y_{i,j} = x_{i,j}$ .

## 4.4 High-pass filters (edge enhancement)

Edges and textures in an image are typical examples of high-frequency information. High-pass filters remove low-frequency image information and therefore enhance high-frequency information such

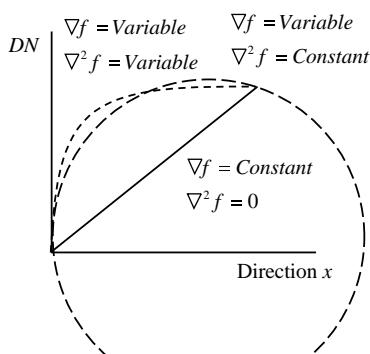
as edges. Most commonly used edge enhancement filters are based on first and second derivatives or the gradient and Laplacian. Given an image  $f(x, y)$ ,

$$\text{Gradient: } \nabla f = \frac{\partial f(x, y)}{\partial x} \bar{i} + \frac{\partial f(x, y)}{\partial y} \bar{j} \quad (4.12)$$

where  $\bar{i}$  and  $\bar{j}$  are unit vectors in the  $x$  and  $y$  directions;

$$\text{Laplacian: } \nabla^2 f = \frac{\partial^2 f(x, y)}{\partial x^2} + \frac{\partial^2 f(x, y)}{\partial y^2}. \quad (4.13)$$

It is important to note that the two types of high-pass filters work in different ways. The gradient is the first derivative at pixel  $f(x, y)$  and as a measurement of DN change rate, it is a vector characterizing the maximum magnitude and direction of the DN slope around the pixel  $f(x, y)$ . The Laplacian, as the second derivative at pixel  $f(x, y)$ , is a scalar that measures the rate of change in gradient. In plain words, the Laplacian describes the curvature of a slope but not its magnitude and direction (discussed again in Section 16.4.2). As shown in Figure 4.9, a flat DN slope has a constant gradient but zero Laplacian because the change rate of a flat slope is zero. For a slope with a constant curvature (an arc of a circle), the gradient is a variable while the Laplacian is a constant. Only for a slope with varying curvature are both gradient and Laplacian variables. This is why the Laplacian suppresses all the image features except sharp boundaries where



**Figure 4.9** Geometric meaning of first and second derivatives

DN gradient changes dramatically, while gradient filtering retains boundary as well as slope information.

An output image of high-pass filtering is usually no longer in the 8 bit positive integer range and so must be rescaled based on the actual limits of the image to 0–255 for display.

#### 4.4.1 Gradient filters

The numerical calculation of gradient based on Equation 4.12 is a simple differencing between a pixel under filtering and its neighbour pixels divided by the distance in between. Gradient filters are always in pairs to produce the  $x$  component ( $g_x$ ) and  $y$  component ( $g_y$ ) or components in diagonal directions:

$$g_x = \frac{f(x, y) - f(x + \delta x)}{\delta x} \quad g_y = \frac{f(x, y) - f(x, y + \delta y)}{\delta y}. \quad (4.14)$$

For instance, the gradient between a pixel  $f(x, y)$  and its next neighbour on the right  $f(x + 1, y)$  is  $g_x = f(x, y) - f(x + 1, y)$ ; here the increment for raster data is  $\delta x = 1$  (Figure 4.10a). We present several of the most commonly used gradient filter kernels which are based on the principles of the equations in (4.14), assuming an odd-number kernel size.

*Gradient filters:*

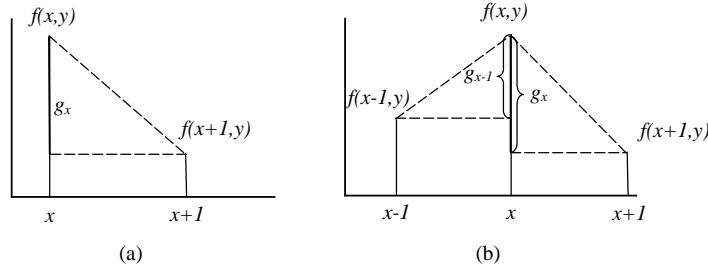
$$g_x = (0 \quad -1 \quad 1) \quad g_y = \begin{pmatrix} 0 \\ -1 \\ 1 \end{pmatrix}.$$

*Prewitt filters:*

$$\begin{pmatrix} -1 & 0 & 1 \\ -1 & 0 & 1 \\ -1 & 0 & 1 \end{pmatrix} \quad \begin{pmatrix} -1 & -1 & -1 \\ 0 & 0 & 0 \\ 1 & 1 & 1 \end{pmatrix}$$

or

$$\begin{pmatrix} -1 & -1 & 0 \\ -1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix} \quad \begin{pmatrix} 0 & 1 & 1 \\ -1 & 0 & 1 \\ -1 & -1 & 0 \end{pmatrix}$$



**Figure 4.10** (a) Calculation of gradient in  $x$  direction  $g_x$ ; and (b) calculation of the  $x$  component of the Laplacian,  $\nabla^2 f_x$

*Sobel filters:*

$$\begin{pmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{pmatrix} \quad \begin{pmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{pmatrix}$$

or

$$\begin{pmatrix} -1 & -2 & 0 \\ -2 & 0 & 2 \\ 0 & 2 & 1 \end{pmatrix} \quad \begin{pmatrix} 0 & 2 & 1 \\ -2 & 0 & 2 \\ -1 & -2 & 0 \end{pmatrix}.$$

The magnitude  $g_m$  and orientation  $g_a$  of a gradient can be computed from  $g_x$  and  $g_y$ :

$$g_m = \sqrt{g_x^2 + g_y^2} \quad g_a = \arctan(g_y/g_x). \quad (4.15)$$

If we apply (4.15) to a digital elevation model (DEM),  $g_m$  produces a slope map and  $g_a$  an aspect map of topography (see also Section 16.4.1).

Figure 4.11 illustrates the results of Sobel filters. The  $g_x$  image in (b) enhances the vertical edges

while the  $g_y$  image in (c) enhances horizontal edges. A filtered image is no longer a positive integer image; it is composed of both positive and negative real numbers as gradient can be either positive or negative depending on whether the DN is changing from dark to bright or vice versa.

#### 4.4.2 Laplacian filters

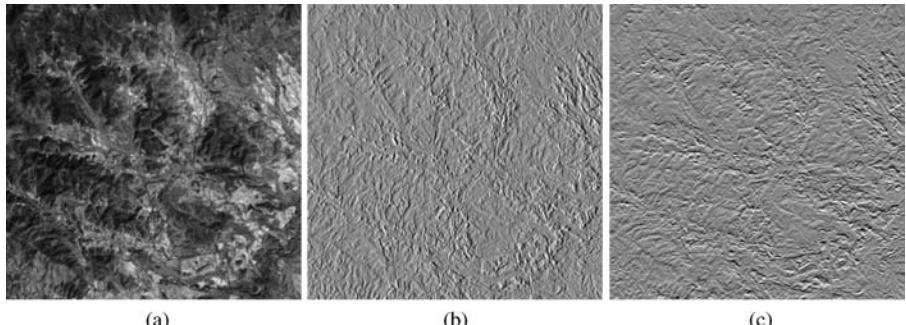
As the second derivative, we can consider the Laplacian as the difference in gradient. Formula (4.13) is composed of two parts: the secondary partial derivative in the  $x$  direction and  $y$  direction. We can then rewrite (4.13) as

$$\nabla^2 f = \nabla^2 f_x + \nabla^2 f_y. \quad (4.16)$$

Let's consider the  $x$  direction as shown in Figure 4.10b:

The gradient at position  $x-1$ :

$$g_{x-1} = f(x-1, y) - f(x, y)$$



**Figure 4.11** Illustration of an image (a) and its  $g_x$  and  $g_y$  for Sobel filters in (b) and (c)

The gradient at position  $x$ :

$$g_x = f(x, y) - f(x + 1, y).$$

Thus we have

$$\nabla^2 f_x = g_{x-1} - g_x = f(x-1, y) + f(x+1, y) - 2f(x, y).$$

Similarly in the  $y$  direction, we have

$$\nabla^2 f_y = g_{y-1} - g_y = f(x, y-1) + f(x, y+1) - 2f(x, y).$$

The above two calculation equations can be translated into a standard Laplacian filter kernel:

$$\begin{pmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{pmatrix}.$$

A more commonly used equivalent form of a Laplacian filter is

$$\begin{pmatrix} 0 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 0 \end{pmatrix}.$$

If we also consider the diagonal directions, then the Laplacian filter is modified as

$$\begin{pmatrix} -1 & -1 & -1 \\ -1 & 8 & -1 \\ -1 & -1 & -1 \end{pmatrix}.$$

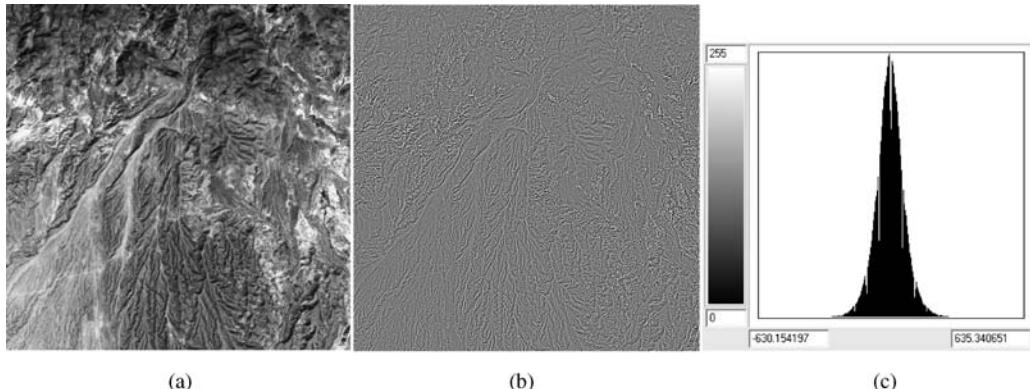
In general, we can consider Laplacian filtering for raster image data as a summation of all the differences between a pixel  $f(x, y)$  and its neighbouring pixels  $f(x + \delta x, y + \delta y)$ :

$$\nabla^2 f = \sum_{\delta x=-1}^1 \sum_{\delta y=-1}^1 [f(x, y) - f(x + \delta x, y + \delta y)]. \quad (4.17)$$

The Laplacian filter produces an image of edges (Figure 4.12). The histogram of such an image is typically symmetrical about a high peak at zero with both positive and negative values (Figure 4.12c). It is important to remember that both very negative and very positive values are edges. As implied by the Laplacian kernels, if the DN of the central pixel in the Laplacian kernel is higher than those of its neighbouring pixels, the Laplacian is positive, indicating a convex edge; otherwise, if the central pixel in the Laplacian kernel is lower than those of its neighbour pixels, the Laplacian is negative, indicating a concave edge.

#### 4.4.3 Edge-sharpening filters

Increasing the central weight of the Laplacian filter by  $k$  is equivalent to adding  $k$  times the original image back to the Laplacian filtered image. The resultant image is similar to the original image but with sharpened edges. The commonly



**Figure 4.12** Illustration of an image (a) and its Laplacian filtering result (b) together with a histogram of the Laplacian image (c)

used add-back Laplacian filters (also called edge-sharpening filters) are

$$\begin{pmatrix} 0 & -1 & 0 \\ -1 & 10 & -1 \\ 0 & -1 & 0 \end{pmatrix} \quad \begin{pmatrix} -1 & -1 & -1 \\ -1 & 14 & -1 \\ -1 & -1 & -1 \end{pmatrix}.$$

The central weight can be changed arbitrarily to control the proportion between the original image and the edge image. This simple technique is popular not only for remote sensing imagery, but also for commercial digital picture enhancement of photographic products.

## 4.5 Local contrast enhancement

The PSF  $h(x, y)$  can be dynamically defined by the local statistics. In this case,  $h(x, y)$  is no longer a predefined fixed function as in the cases of smoothing, gradient and Laplacian filters. It varies with  $f(x, y)$  according to image local statistics. This branch of filtering techniques is called adaptive filtering. Adaptive filtering can be used for contrast enhancement, edge enhancement and edge-preserving smoothing as well as noise removal.

One typical adaptive filter is the local contrast enhancement. The purpose of local contrast enhancement is to produce the same contrast in every local region throughout an image. An adaptive algorithm is employed to adjust parameters of a point operation function pixel by pixel, based on local statistics, so as to achieve contrast enhancement. This technique represents the combination of a point operation and neighbourhood processing. We therefore introduce it in this chapter rather than in Chapter 2 on point operations.

The simplest local contrast enhancement is the local mean adjustment technique. Here we use the same notation as in Chapter 2 to illustrate better the processing as a contrast enhancement. Let  $\bar{x}_{ij}$  be the local mean in some neighbourhood of pixel  $x_{ij}$ , say a  $31 \times 31$  square window centred at  $(i, j)$ . Then

$$y_{ij} = x_{ij} + m_o - \bar{x}_{ij}. \quad (4.18)$$

This technique adjusts local brightness to the global mean  $m_o$  of the image while leaving the local contrast unchanged. For the whole image, the processing may reduce the image global contrast

(standard deviation) but will maintain the average brightness (global mean).

Let  $\bar{x}_{ij}$  and  $\sigma_{ij}$  be the local mean and local standard deviation in some neighbourhood of pixel  $x_{ij}$ ; then a local contrast enhancement algorithm using a linear function is defined as follows:

$$y_{ij} = \bar{x}_{ij} + (x_{ij} - \bar{x}_{ij}) \frac{\sigma_o}{\sigma_{ij} + 1} \quad (4.19)$$

where the 1 in the denominator is to prevent overflow when  $\sigma_{ij}$  is almost 0.

This local enhancement function stretches  $x(i, j)$  to achieve a predefined local standard deviation  $\sigma_o$ . In addition, the local mean can also be adjusted by modifying (4.19) as

$$y_{ij} = \alpha m_o + (1 - \alpha) \bar{x}_{ij} + (x_{ij} - \bar{x}_{ij}) \frac{\sigma_o}{\sigma_{ij} + 1} \quad (4.20)$$

where  $m_o$  is the mean to be enforced locally and  $0 \leq \alpha \leq 1$  is a parameter to control the degree to which it is enforced.

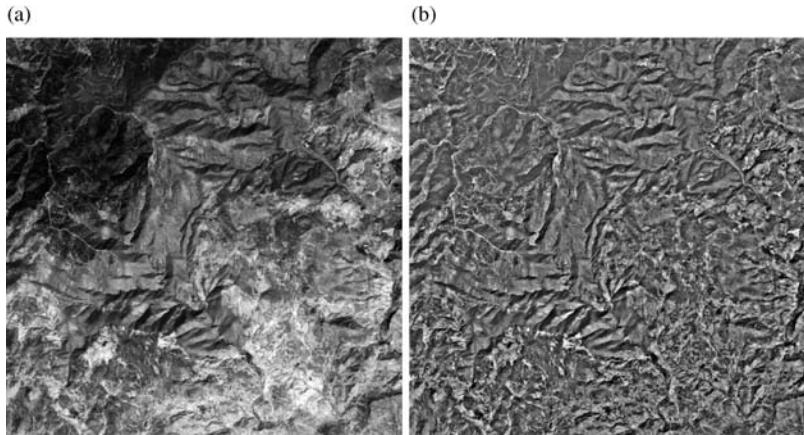
The function defined by (4.20) will produce an image with a local mean  $m_o$  and local standard deviation  $\sigma_o$  everywhere in the image. The actual local mean  $m_o$  and local standard deviation  $\sigma_o$  vary from pixel to pixel in a certain range depending on the strength of parameters  $\alpha$ .

It is important to keep in mind that local contrast enhancement is not a point operation. It is essentially a neighbourhood processing. This technique may well enhance localized subtle details in an image but it will not preserve the original image information (Figure 4.13).

## 4.6 \*FFT selective and adaptive filtering

This section presents our recent research illustrating how filters are designed based on the spatial pattern of targeted periodic noise to be removed (Liu and Morgan, 2006).

Remotely sensed images or products derived from these images can be contaminated by systematic noise of a particular frequency or frequencies which vary according to some function relating to the sensor or imaging configuration. To remove this type of noise pattern, FFT filtering in the frequency domain is the most effective approach.



**Figure 4.13** Effects of local enhancement: (a) original image; and (b) local enhancement image

As shown in Figure 4.14a, the image is a feature-matching result of a pair of Landsat-7 ETM+ panchromatic images across a major earthquake event, aiming to reveal the co-seismic shift at sub-pixel accuracy (Liu, Mason and Ma, 2006). Severe horizontal and vertical striping noise patterns plague the image and seriously obscure the desired surface shift information.

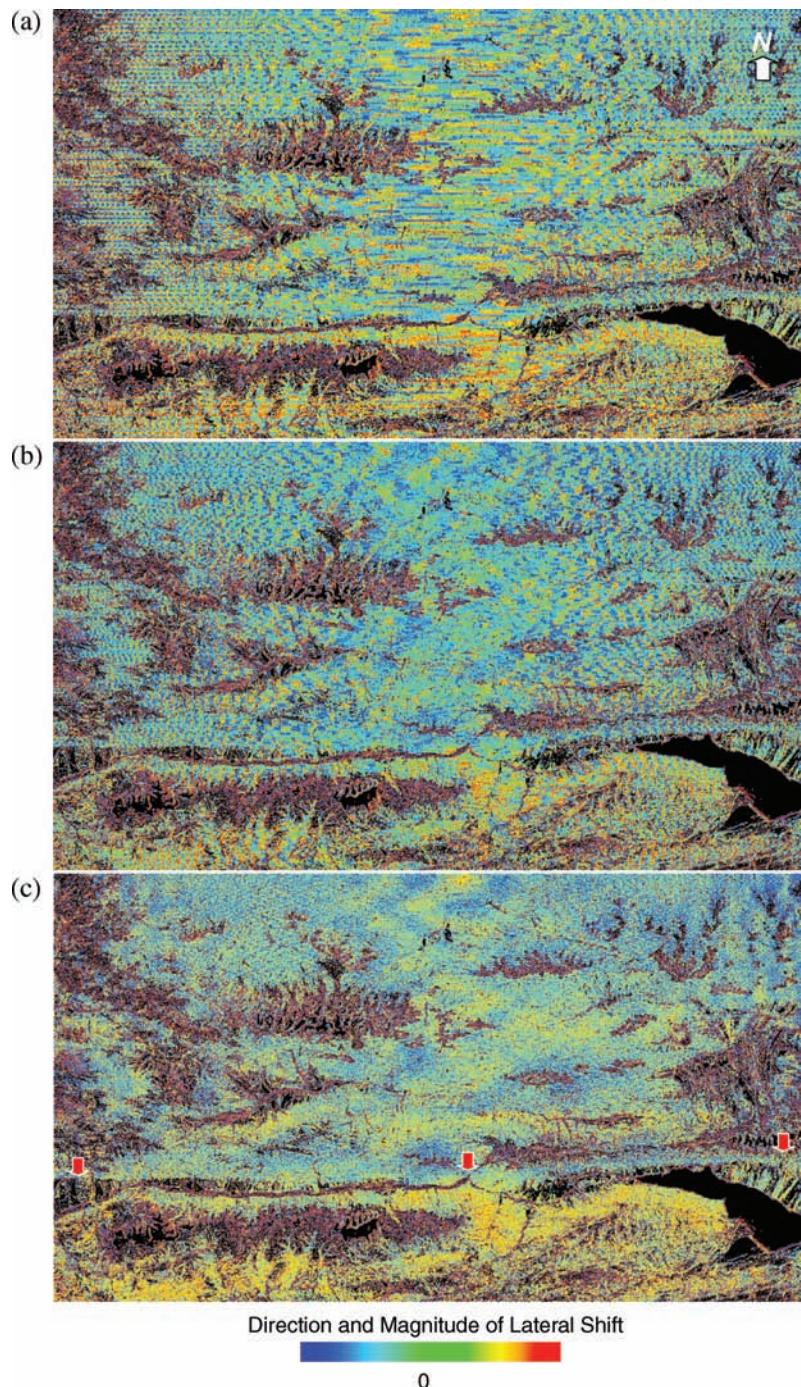
#### 4.6.1 FFT selective filtering

In-depth investigation indicates that the horizontal striping is caused by the two-way scanning of the Landsat ETM+ in conjunction with the orbital drift between two image acquisitions. Figure 4.15 left shows a small sub-scene of an area of the image which is relatively homogeneous. In general, the clear striping pattern is representative of the entire image, displaying equally spaced lighter and darker bands. This regular noise pattern was significantly enhanced by applying a 1 line  $\times$  101 column smoothing filter to remove much of the scene content (Figure 4.15 right). This striping noise of fixed frequency can be removed by frequency domain filtering. After converting the image  $f(x, y)$  into a frequency spectrum  $F(u, v)$  via the FFT, the key point for successful filtering is to locate the representative frequency spectrum ‘spots’ corresponding to the periodic noise pattern and mask them off with a function  $H(u, v)$  before making the inverse FFT back to an image. Such FFT selective

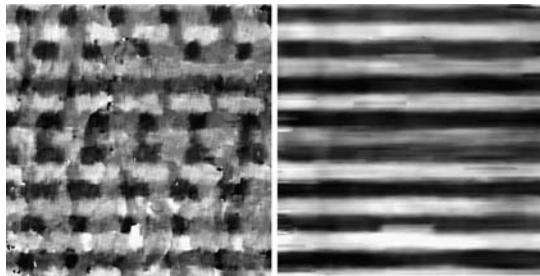
filtering comprises the following steps in two stages (Figure 4.16):

Procedure of FFT selective filtering:

1. A single sub-scene (e.g.  $1024 \times 1024$  pixels) of the image is selected from an area with relatively homogeneous scene content and clear striping.
2. The stripes are enhanced by a  $1 \times 101$  one-dimensional horizontal smoothing filter.
3. The stripe-enhanced sub-scene is then thresholded to create a black and white binary image.
4. The binary stripe image is transformed to the frequency domain, using the 2D FFT, and the power spectrum is calculated to give the absolute magnitude of the frequency components.
5. The magnitude of the spectrum is normalized to 0–1 after masking off the dominant zero and very near-zero frequencies that always form the highest magnitudes in the frequency spectrum but which carry no periodic noise information. The frequencies of the highest magnitude peaks which relate to specific noise components can then be located and retained by thresholding.
6. Inverting this result (i.e. interchanging zeros for ones and ones for zeros) creates an instant binary mask for selective frequency filtering. The zero and very near-zero frequencies are set to 1 in the mask to retain the natural scene content in filtering. Additionally, this initial binary filter function is convolved with an



**Figure 4.14** (a) The original co-seismic shift image. (b) The FFT selective filtering result; the horizontal stripes have been successfully removed, but the vertical noise is now very noticeable. (c) The FFT adaptive filtering result; the multiple frequency wavy patterns of vertical stripes have been successfully removed and thus clearly reveal the phenomena of regional co-seismic shift along the Kunlun fault line as indicated by three arrows. Yellow-red indicates movement to the right and cyan-blue indicates movement to the left



**Figure 4.15** A  $500 \times 500$  pixel sub-scene of the image (left); and (right) the image filtered with a one-dimensional, horizontal smoothing filter (kernel size  $1 \times 101$ ) to isolate the noise and aid analysis

- appropriately sized Gaussian pulse to eliminate the possibility of ringing artefacts.
7. The mask generated from the selected sub-scene can then be used on the entire noisy image, because the identified striping frequencies are independent of position in the image. The mask can then simply be multiplied by the 2D FFT of the entire image.
  8. Finally, the result is transformed back to the image domain, via the 2D inverse FFT, to produce a de-striped image.

#### 4.6.2 FFT adaptive filtering

As shown in Figures 4.17 and 4.14b, after the successful removal of the horizontal stripes, the vertical noise pattern in the horizontally filtered image is more clearly revealed than before. The noise is not

simple stripe-like but instead forms a series of parallel, vertically aligned wavy lines with progressively increasing frequency from the middle to both edges of the image (Figure 4.14b). Analysis and simulation of the changing frequency of the noise indicate that this is caused by the transition from the fixed angular speed ( $2.210\ 95\ \text{rad s}^{-1}$ ) of the ETM+ scanner to the line speed on the curved Earth's surface, as depicted in Figure 4.18. The surface scanning speed, with a fixed angular speed of the rotating scanner mirror, varies symmetrically across the swath width about the nadir point, accelerating out towards the edges as defined by the formula (4.21) derived from the diagram in Figure 4.18.

The scanning speed on the Earth's surface,  $dl/dt$ , varies as a function of  $\theta$  or the scanning angular speed  $d\theta/dt$ . By the sine theorem

$$\frac{R}{\sin \theta} = \frac{R+r}{\sin \alpha}.$$

Then,

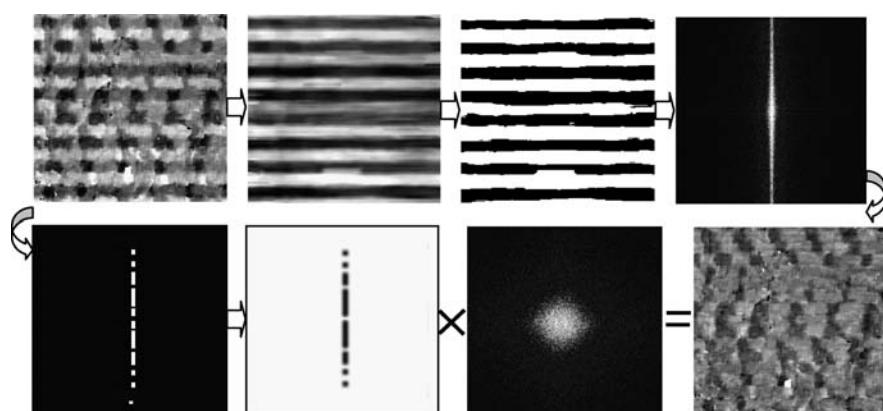
$$\alpha = \arcsin\left(\frac{R+r}{R} \sin \theta\right).$$

As  $\varphi = \pi - \alpha - \theta$ , then

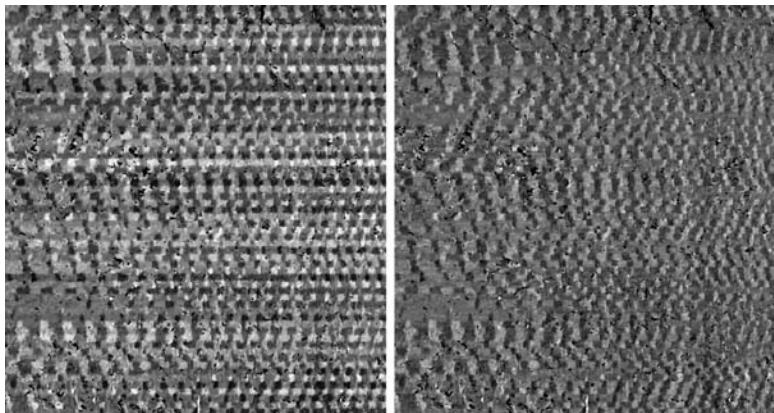
$$\varphi = \pi - \theta - \arcsin\left(\frac{R+r}{R} \sin \theta\right).$$

Given the arc length  $l = R\varphi$ , then

$$\frac{dl}{dt} = R \frac{d\varphi}{dt} = R \frac{d\varphi}{d\theta} \cdot \frac{d\theta}{dt}.$$



**Figure 4.16** A graphical illustration, step by step, of the selective frequency filtering algorithm for the removal of horizontal stripes in the image



**Figure 4.17** After the selective horizontal filtering applied to the original image (left) to remove horizontal stripes successfully, the vertical wavy stripe noise becomes more obvious in the filtered image (right). Image size:  $2000 \times 2000$

Thus

$$\frac{dl}{dt} = -R \frac{d\theta}{dt} \left[ 1 + \frac{(R+r)\cos\theta}{R\sqrt{1-\left(\frac{R+r}{R}\sin\theta\right)^2}} \right] \quad (4.21)$$

where,

$\theta$  is the scanning angle (variable)

$\varphi$  is the angle subtended by two Earth radii, corresponding to  $\theta$

$\alpha$  is the angle between the scanning line of sight and the Earth's radius at the scanning position

$r$  is the height of the sensor above the Earth's surface

$R$  is the radius of the Earth

$l$  is the scanning length on Earth's surface for angle  $\theta$  from nadir to scanning position, equivalent to the arc length between Earth's radii.

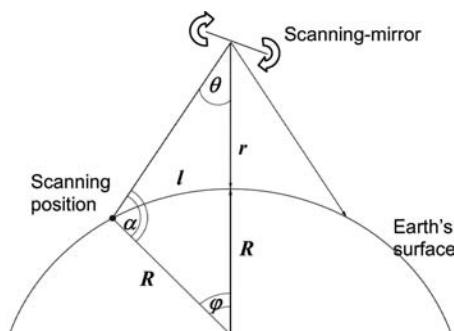
Within the scanning range of the ETM+ sensor system, the function defined by (4.21) can be precisely simulated by the least squares fit of a second-order polynomial, resulting in a parabolic curve (Figure 4.19):

$$f(x) = 4.586 \times 10^{-10}x^2 - 1.368 \times 10^{-7}x + 0.0019 \quad (4.22)$$

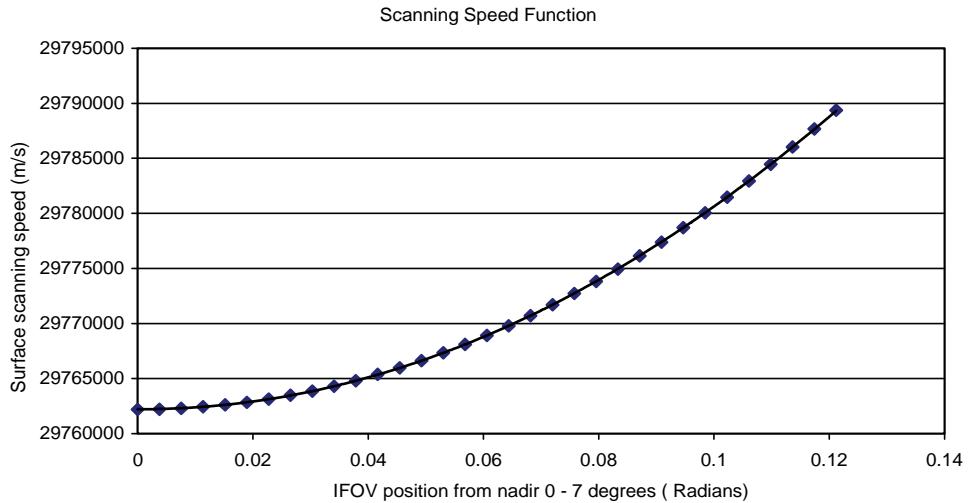
where  $x$  is the image column position.

For processing efficiency, this simple parabolic function is used to adapt the filter aperture in the frequency domain to remove vertical noise patterns of differing frequencies at different image columns. The idea is that each image column should be filtered according to a noise frequency calculated from (4.22) at the corresponding scanning angle  $\theta$ , so the image will be adaptively filtered, column by column, with decreasing frequency from the left to the middle, and then increasing frequency from the middle to the right.

In practice, the wavy form of the stripes makes this a truly 2D filtering problem, in which frequency components will be located not only on the horizontal frequency axis but also at locations diagonal to the spectrum centre. The adaptive filter design must therefore be based on the



**Figure 4.18** The geometric relationship between a constant angular speed of cross-track scanning and the corresponding line speed on the curved surface of the Earth



**Figure 4.19** A plot of the scanning speed function of the Earth's surface as defined by the formula (4.21) in the scanning angular range 0–7° (in diamond markers) fitted by the parabolic curve defined in formula (4.22)

concept of a circularly symmetric band-reject filter of the Butterworth type. The transfer function for a Butterworth band-reject filter of order  $n$  is defined as

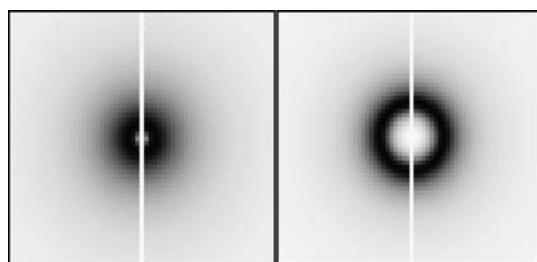
$$H(u, v) = \frac{1}{1 + \left[ \frac{D(u, v)W}{D^2(u, v) - D_0^2} \right]^{2n}} \quad (4.23)$$

where  $D(u, v)$  is the distance from the point  $(u, v)$  to the centre origin of the frequency spectrum,  $D_0$  is the radius of the filter defined as the principal frequency to be filtered out and  $n$  is the order of the filter roll-off either side of the central ring defined by  $D_0$  forming the bandwidth  $W$ .

For an  $M \times N$  image section and its subsequent FT,  $D(u, v)$  is given by

$$D(u, v) = [(u - M/2)^2 + (v - N/2)^2]^{1/2}. \quad (4.24)$$

The filter function,  $H(u, v)$ , thus removes a band of frequencies at a radial distance,  $D_0$ , from the frequency spectrum centre while smoothly attenuating the frequencies either side of  $D_0$  to reduce the possibility of ringing. The key aspect of this adaptive filter design is that  $D_0$  is decided by the scanning function formula (4.21) or (4.22). Thus  $D_0$  varies column by column given the scanning angle  $\theta$  corresponding to the image column position (Figure 4.20).



**Figure 4.20** Two  $512 \times 512$  adaptive Butterworth band-reject filters defined by the function (4.23): left,  $D_0 = 3$ ,  $W = 3$ ,  $n = 3$ ; and right,  $D_0 = 8$ ,  $W = 3$ ,  $n = 3$ . The filters have been edited to allow the central vertical frequencies through without suppression because these components have undergone previous horizontal filtering. The radius of the filter  $D_0$  varies with the image column position according to the formulae (4.22)

The FFT frequency-adaptive filtering procedure is as follows:

1. Starting from the left edge of the image, the FFT for a  $512 \times 512$  neighbourhood is applied to produce  $F(u, v)$ .
2. The noise frequency at the central column of the  $512 \times 512$  neighbourhood is calculated using formula (4.22), given the column position.
3. In the Fourier domain, the noise frequency in  $F(u, v)$  is masked off by multiplication with the filter transfer function  $H(u, v)$  to produce  $F(u, v)H(u, v)$ .
4. Use the IFFT to transform the filtered frequency spectrum  $F(u, v)H(u, v)$  back to an image  $f(x, y) * h(x, v)$ , but only retaining the central column.
5. Move to the next column and repeat steps 1–4 till the last column is reached.
6. Move down to the next  $512$  block below and repeat steps 1–5 till the end of the image is reached.

After the FFT selective and adaptive filtering the horizontal and vertical noise patterns are effectively removed, as shown in Figure 4.14c, and the left lateral co-seismic shift along the Kunlun fault line is clearly revealed.

## 4.7 Summary

Image filtering is a process for removing image information of particular frequencies. Within this context, it is typical of signal processing in the frequency domain via the Fourier transform. On the basis of the convolution theorem, FT-based filtering can be performed in the image domain, using convolution and realized by neighbourhood processing in an image. The most commonly used filters for digital image filtering rely on the concept of convolution and operate in the image domain for reasons of simplicity and processing efficiency.

Low-pass filters are mainly used to smooth image features and to remove noise but often at the cost of degrading image spatial resolution (blurring). To remove random noise with the minimum degradation of resolution, various edge-preserved filters have been developed such as the adaptive median filter. A classification image is a ‘symbol’ image

rather than a digital image and therefore should not be subject to any numerical operations. The mode (majority) filter is suitable for smoothing a classification image as the filtering process is based on the election of a local majority within the processing window without numerical operations.

There are two different types of high-pass filters: gradient and Laplacian filters. As the first derivative of DN change in a direction, the gradient gives a measurement of DN slope. Gradient is a vector and so gradient filters are directional; they are commonly used as orthogonal pairs for directional enhancement. Images representing the magnitude and orientation of gradient can be calculated from the pair of images derived from the orthogonal filters. Laplacian, as the second derivative, is a scalar that measures the change rate of DN slope. Image edge features are characterized as significant DN slope changes and Laplacian is therefore very effective for enhancing and extracting them. One of the most common applications of Laplacian for not only remote sensing image processing but also general graphic enhancement is the so-called edge-sharpening filter.

Combining neighbourhood processing with point operations for contrast enhancement formulates a new method of image processing: local contrast enhancement. It adjusts the image contrast based on local statistics calculated in a neighbourhood. This is based on the concept of a more general branch of neighbourhood processing: adaptive filters. As examples, image-characteristics-based derivation of FFT selective and adaptive filters is introduced for advanced readers at the end of this chapter.

Our general advice on all filtering and neighbourhood processing is not to trust them blindly! Artefacts can be introduced, so use of the original images as a reference is always recommended.

## Questions

- 4.1 Using a diagram, illustrate the 4/f optical image filtering system and explain the principle of image filtering based on the Fourier transform.
- 4.2 What is the convolution theorem and why is it important in digital image filtering?
- 4.3 Explain the relationship between the filtering function  $H(u, v)$  in the frequency domain and

- the PSF function  $h(x, y)$  in the spatial (image) domain.
- 4.4 If the range over which the PSF  $h(x, y)$  is non-zero is  $(-w, +w)$  in one dimension and  $(-t, +t)$  in the other, write down the discrete form of convolution  $f(x, y) * h(x, y)$ .
- 4.5 What is a low-pass filter for digital image filtering and what are its effects? Give some examples of low-pass filter kernels.
- 4.6 Discuss the major drawback of mean filters and the importance of edge-preserved smoothing filters.
- 4.7 To smooth a classification image, what filter is appropriate and why? Describe this filter with an example.
- 4.8 Give a general definition of conditional filters.
- 4.9 Describe the clean pixels filter and explain how it works.
- 4.10 Describe the  $k$  nearest mean filter, median filter and adaptive median filter and discuss their merits based on the filtering results of the sample image below:
- 4.11 What is it meant by high-pass filtering?
- 4.12 Describe the mathematical definitions of image gradient and Laplacian together with examples of gradient and Laplacian filters.
- 4.13 Use a diagram to illustrate and explain the different functionalities of gradient- and Laplacian-based high-pass filters.
- 4.14 Given a DEM, how would you calculate the slope and aspect of topography using gradient filters?
- 4.15 Why is the histogram of a Laplacian filtered image symmetrical about a high peak at zero with both positive and negative values?
- 4.16 What is an edge-sharpening filter? What are the major applications of edge- sharpening filters?
- 4.17 Describe local contrast enhancement technique as a neighbourhood processing procedure and explain why it is not a point operation.

					Median filter	Adapt median filter	$K$ nearest mean filter
173	140	124	113	100			
167	145	136	18	83			
138	252	122	96	117			
144	134	83	87	116			
137	115	95	119	142			



# 5

## RGB–IHS Transformation

In this chapter, we first describe the principles of the RGB–IHS and IHS–RGB transformations. Two decorrelation stretch techniques, both based on saturation stretch, are then discussed. Finally, a hue RGB (HRGB) colour composition technique is introduced. The RGB–IHS transformation is also a powerful tool for data fusion but we leave this part to Chapter 6 along with the discussion of several other data fusion techniques.

### 5.1 Colour coordinate transformation

A colour is expressed as a composite of three primaries, *Red*, *Green* and *Blue*, according to the tristimulus theory. For colour perception on the other hand, a colour is quantitatively described in terms of three variables, *Intensity*, *Hue* and *Saturation*, which are measurements of the brightness, spectral range and purity of a colour. There are several variants of the RGB–IHS transformation based on different models. For the RGB additive colour display of digital images, a simple RGB colour cube is the most appropriate model. The RGB–IHS colour coordinate transformation in a colour cube is similar to a three-dimensional Cartesian–conical coordinate transformation.

As shown in Figure 5.1, any a colour in a three-band colour composite is a vector  $\mathbf{P}(r, g, b)$  within

a colour cube of 0–255 in three dimensions (for 24 bit RGB colour display). The major diagonal line connecting the origin and the furthest vertex is called the *grey line* because the pixels lying on this line have equal components in red, green and blue ( $r = g = b$ ). The intensity of a colour vector  $\mathbf{P}$  is defined as the length of its projection on the grey line, OD, the hue as the azimuth angle around the grey line,  $\alpha$ , and the saturation as the angle between the colour vector  $\mathbf{P}$  and the grey line,  $\varphi$ . Let the hue angle of pure blue colour be zero. We then have the following RGB–IHS transformation:

$$I(r, g, b) = \frac{1}{\sqrt{3}}(r + g + b) \quad (5.1)$$

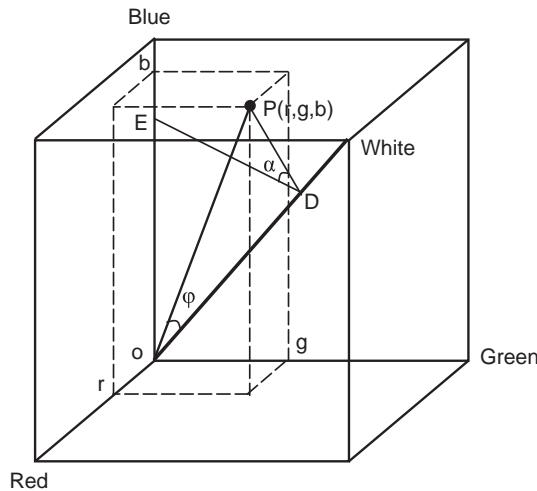
$$H(r, g, b) = \arccos \frac{2b - g - r}{2V}$$

where  $V = \sqrt{(r^2 + g^2 + b^2) - (rg + rb + gb)}$  (5.2)

$$S(r, g, b) = \arccos \frac{r + g + b}{\sqrt{3(r^2 + g^2 + b^2)}}. \quad (5.3)$$

Saturation as defined in formula (5.3) can then be rewritten as a function of intensity:

$$S(r, g, b) = \arccos \frac{I(r, g, b)}{\sqrt{r^2 + g^2 + b^2}}. \quad (5.4)$$



**Figure 5.1** The colour cube model for the RGB–IHS transformation

For the same intensity, if  $r=g=b$ , saturation reaches its minimum,  $S(r, g, b) = \arccos 1 = 0$ ; if two of  $r, g, b$  are equal to 0, then saturation reaches its maximum,  $S(r, g, b) = \arccos(1/\sqrt{3}) \approx 54.7356^\circ$ . Actually, saturation of a colour is the ratio between its achromatic and chromatic components, so saturation increases with the increase of difference between  $r, g$  and  $b$ , and it can therefore be defined by the maximum and minimum of  $r, g$  and  $b$  in a value range from no saturation (0) to full saturation (1) as in the formula below (Smith, 1978):

$$S(r, g, b) = \frac{\max(r, g, b) - \min(r, g, b)}{\max(r, g, b)}. \quad (5.5)$$

This formula implies that a colour vector reaches full saturation if at least one of its  $r, g$  and  $b$  components is equal to 0 while not all of them are 0. For instance, colour  $\mathbf{P}(r, g, b) = (255, 0, 0)$  is pure red with full saturation and  $\mathbf{P}(r, g, b) = (0, 200, 150)$  is a greenish cyan with full saturation.

The value range of hue is  $0\text{--}2\pi$  or  $0^\circ\text{--}360^\circ$ , while the value range of the arccosine function of hue in formula (5.2) is  $0\text{--}\pi$ , but the  $2\pi$  range of hue can be determined based on the relationship between  $r, g$  and  $b$ . For instance, if  $b > r > g$ , the actual hue angle is  $\text{hue}(r, g, b) = 2\pi - H(r, g, b)$ .

Given intensity  $I$ , hue angle  $\alpha$  and saturation angle  $\varphi$ , we can also derive the IHS–RGB transfor-

mation based on the same 3D geometry depicted in Figure 5.1:

$$B(I, \alpha, \varphi) = \frac{I}{\sqrt{3}}(1 + \sqrt{2} \tan \varphi \cos \alpha) \quad (5.6)$$

$$G(I, \alpha, \varphi) = \frac{I}{\sqrt{3}} \left[ 1 - \sqrt{2} \tan \varphi \cos \left( \frac{\pi}{3} + \alpha \right) \right] \quad (5.7)$$

$$R(I, \alpha, \varphi) = \frac{I}{\sqrt{3}} \left[ 1 + \sqrt{2} \tan \varphi \cos \left( \frac{2\pi}{3} + \alpha \right) \right]. \quad (5.8)$$

Equivalently, but with a slight difference, the RGB–IHS transformation can also be derived from matrix operations by a coordinate rotation of the colour cube, and aided by sub-coordinates of  $v_1$  and  $v_2$ . As shown in Figure 5.2, the sub-axis  $v_1$  is perpendicular to the grey line starting from the intensity  $I$ ; it is in the plane decided by the blue axis and the grey line. The sub-axis  $v_2$  is perpendicular to both the grey line and  $v_1$ . Thus  $v_1$  and  $v_2$  formulate a plane perpendicular to the grey line and the end point of the colour vector  $\mathbf{P}(r, g, b)$  is in this plane. Thus considering the Cartesian–polar coordinate transformation of the sub-coordinate system of  $v_1$  and  $v_2$ , the following matrix operation between  $I, v_1, v_2$  and  $R, G, B$  can be established:

$$\begin{pmatrix} I \\ v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} 1/3 & 1/3 & 1/3 \\ -1/\sqrt{6} & -1/\sqrt{6} & 2/\sqrt{6} \\ 1/\sqrt{6} & -2/\sqrt{6} & 0 \end{pmatrix} \begin{pmatrix} R \\ G \\ B \end{pmatrix}. \quad (5.9)$$

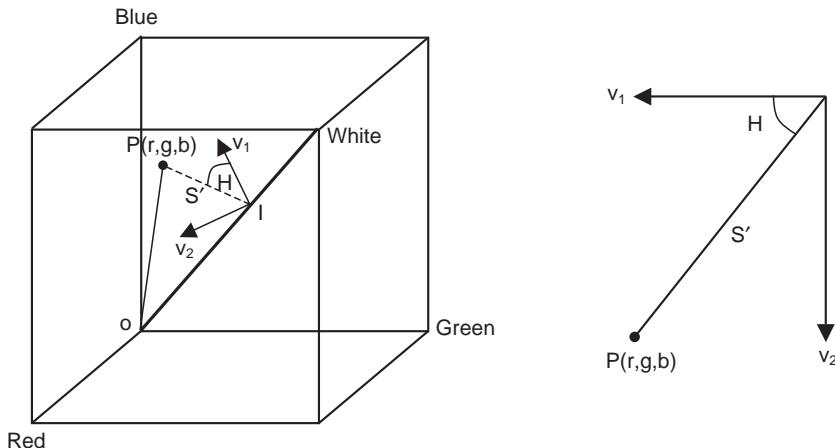
Then the hue and saturation can be derived based on their relationships with  $v_1$  and  $v_2$  (Figure 5.2):

$$H = \arctan(v_2/v_1) \quad (5.10)$$

$$S' = \sqrt{v_1^2 + v_2^2} \quad (5.11)$$

$$S = \arctan \frac{S'}{\sqrt{3}I}. \quad (5.12)$$

Here  $S'$  is the saturation for a given intensity  $I$  while  $S$  is the intensity-scaled angular saturation as depicted in Figure 5.1. Depending on the RGB–IHS model used, there are several different definitions for saturation. Many publications define saturation



**Figure 5.2** The model of the matrix RGB-IHS transformation. Adapted from Niblack (1986)

in the form of (5.11) however, this definition is correct only for a fixed intensity. For a digital image RGB additive colour composite display based on the RGB colour cube model, the definitions of saturation in angle  $\varphi$  given in (5.4) and (5.5) are the most appropriate. Formula (5.12) is essentially the same as (5.4).

An IHS–RGB transformation can then be derived from the inversion of (5.9):

$$\begin{pmatrix} R \\ G \\ B \end{pmatrix} = \begin{pmatrix} 1 & -1/2\sqrt{6} & 3/2\sqrt{6} \\ 1 & -1/2\sqrt{6} & -3/2\sqrt{6} \\ 1 & 1/\sqrt{6} & 0 \end{pmatrix} \begin{pmatrix} I \\ v_1 \\ v_2 \end{pmatrix} \quad (5.13)$$

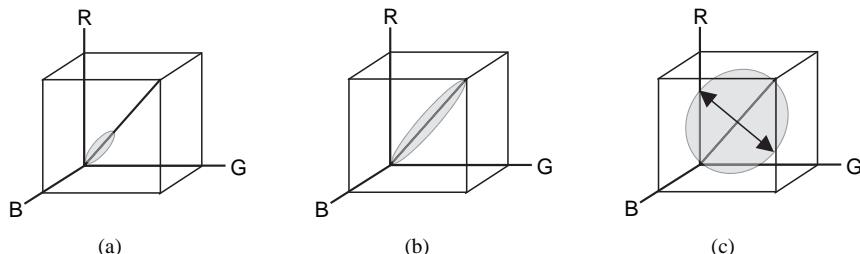
$$v_1 = S' \cos 2\pi H \quad (5.14)$$

$$v_2 = S' \sin 2\pi H.$$

RGB–IHS and IHS–RGB transformations allow us to manipulate intensity, hue and saturation components separately and thus enable some innovative processing for decorrelation stretch and image data fusion.

## 5.2 IHS decorrelation stretch

High correlation generally exists among spectral bands of multi-spectral images. As a result, the original image bands displayed in RGB formulate a slim cluster along the grey line occupying only a very small part of the space of the colour cube (Figure 5.3a). Contrast enhancement on individual image bands can elongate the cluster in the colour cube but it is not effective for increasing the volume of the cluster since it is equivalent to stretching the



**Figure 5.3** (a) Distribution of pixels in the RGB cube for typical correlated bands. (b) The effect of stretching individual bands. The data cluster is elongated along the grey line but not stretched to fill the RGB cube by this operation. (c) Decorrelation stretch expands the data cluster in the direction perpendicular to the grey line to fill the 3D space of the RGB cube

**Table 5.1** The correlation coefficients before and after decorrelation stretch of the Landsat-7 ETM+ bands 5, 3 and 1 RGB colour composite shown in Figure 5.4

Correlation matrix before DS				Correlation matrix after DS			
Correlation	Band 1	Band 3	Band 5	Correlation	Band 1	Band 3	Band 5
Band 1	1.00	0.945	0.760	Band 1	1.00	0.842	0.390
Band 3	0.945	1.00	0.881	Band 3	0.842	1.00	0.695
Band 5	0.760	0.881	1.00	Band 5	0.390	0.695	1.00

intensity only (Figure 5.3b). To increase the volume, the data cluster should expand in both directions along and perpendicular to the grey line. This is equivalent to stretching both intensity and saturation components (Figure 5.3c). The processing is called *IHS decorrelation stretch (IHSDS)* because the correlation among the three bands is reduced to generate a spherical data cluster in the RGB cube as indicated in Table 5.1. In comparison with an ordinary contrast stretch, the IHSDS is essentially a saturation stretch. As proposed by Gillespie, Kahle and Walker (1986), the IHSDS technique involves the following steps:

1. RGB–IHS transformation.
2. Stretch intensity  $I$  and saturation  $S$  components.
3. IHS–RGB transformation.

In the second step, the hue component can also be stretched. However, when transforming back to RGB display, the resultant colours may not be comparable with those of the original image and this makes image interpretation potentially difficult.

The limited hue range of a colour composite image is mainly caused by colour bias. If the average brightness of one band is significantly higher than those of the other two bands, the colour composite will have an obvious colour ‘cast’ of the primary colour assigned to the band of highest intensity. As we discussed in Chapter 2, the balanced contrast enhancement technique (BCET) was developed to solve just this problem. BCET removes inter-band colour bias and therefore increases the hue variation of a colour composite. As a result, the hue component derived from a BCET stretched colour composite will have a much wider value range than that derived from the original colour-biased composite.

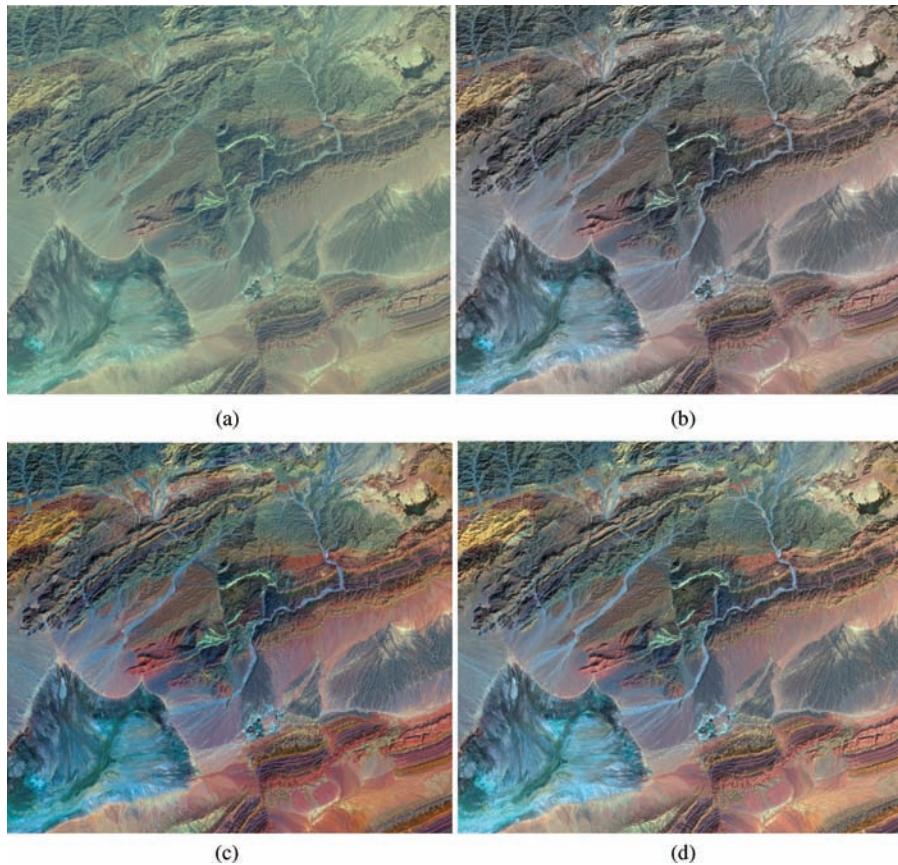
The wider hue value range achieved by BCET means more spectral information is presented as hue rather than as intensity and saturation, which are fundamentally different from the wide hue range which would be achieved by stretching the hue component. In many cases, a simple linear stretch with automatic clipping or an interactive piecewise linear stretch can also effectively eliminate the colour bias. An optimized IHSDS can be achieved by performing BCET or linear stretch as a pre-processing step, as summarized below:

1. BCET stretch (or linear stretch with appropriate clipping).
2. RGB–IHS transformation.
3. Saturation component stretching.
4. IHS–RGB transformation.

The DN ranges of the images converted from stretched IHS components back to RGB coordinates may exceed the maximum range of display device (usually 8 bits or 0–255 per channel) and so may need to be adjusted to fit the maximum 8 bit DN range. This can be done easily in any image processing system, such as ER Mapper; the image will be automatically displayed within 8 bits per channel using the actual limits of input image DNs.

The effect of the IHSDS-based saturation stretch is similar to that of the decorrelation stretch based on principal component analysis (to be introduced in Chapter 7). The difference between them is that principal component decorrelation stretch is based on scene statistics while the IHS decorrelation stretch is interactive, flexible and based on user observation of the saturation image and its histogram.

Decorrelation stretch enhances the colour saturation of a colour composite image and thus



**Figure 5.4** Colour composites of ETM+ bands 5, 3 and 1 in red, green and blue: (a) Original colour composite without any stretch; (b) BCET stretched colour composite; (c) IHS decorrelation stretched colour composite after BCET stretch; and (d) DDS ( $k = 0.5$ ) colour composite after BCET stretch

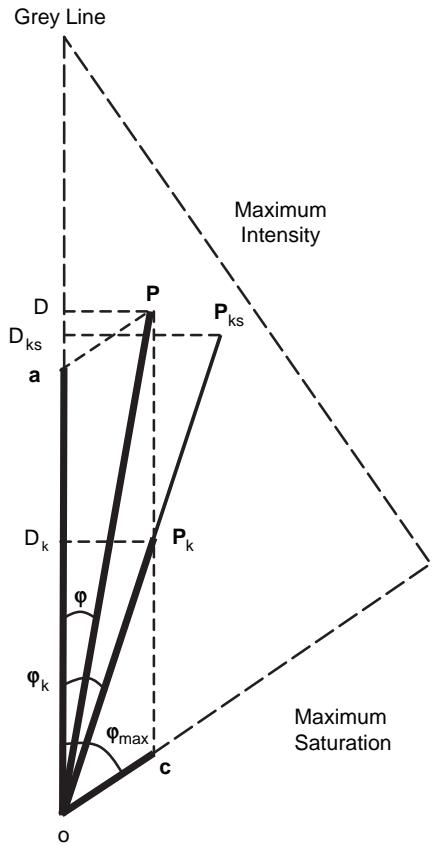
effectively improves the visual quality of the image spectral information, without significant distortion of its spectral characteristics as illustrated in Figure 5.4. Decorrelation-stretch-enhanced colour images are easy to understand and interpret and have been successfully used for many applications of remote sensing.

### 5.3 Direct decorrelation stretch technique

This technique performs a *direct saturation stretch* (DDS) without using RGB–IHS and IHS–RGB transformations (Liu and Moore, 1996). The DDS achieves the same effect as the IHSDS. As DDS involves only simple arithmetic operations and

can be controlled quantitatively, it is much faster, more flexible and more effective than the IHSDS technique.

As shown in Figure 5.1, a colour vector,  $\mathbf{P}$ , and the grey line together define a plane or a slice of the RGB cube. If we take this slice out as shown in Figure 5.5, the grey line, the full saturation line and the maximum intensity line formulate a triangle that includes all the colours with the same hue but various intensity and saturation. The colour vector  $\mathbf{P}$  is between the grey (achromatic) line and the maximum saturation (chromatic) line and it can be considered as the sum of two vectors: a vector  $\mathbf{a}$  representing the achromatic (zero-saturation) component, the white light in the colour, and a vector  $\mathbf{c}$  representing the chromatic (full-saturation) component that is relevant to the pure colour of the hue.



**Figure 5.5** The principle of the direct decorrelation stretch (DDS) technique

Given  $\mathbf{P} = (r, g, b)$ , let  $a = \min(r, g, b)$ . Then

$$\begin{aligned}\mathbf{a} &= (a, a, a) \\ \mathbf{c} &= (r - a, g - a, b - a) \quad (5.15) \\ &= \mathbf{P} - \mathbf{a}\end{aligned}$$

or

$$\mathbf{P} = \mathbf{a} + \mathbf{c}. \quad (5.16)$$

A DDS is achieved by reducing the achromatic component  $\mathbf{a}$  of the colour vector  $\mathbf{P}$ , as defined below:

$$\mathbf{P}_k = \mathbf{P} - k\mathbf{a} \quad (5.17)$$

where  $k$  is an achromatic factor and  $0 < k < 1$ .

As shown in Figure 5.5, the operation shifts the colour vector  $\mathbf{P}$  away from the achromatic line to form a new colour vector  $\mathbf{P}_k$  with increased saturation ( $\varphi_k > \varphi$ ) and decreased intensity ( $OD_k < OD$ ). To

restore the intensity to a desired level, linear stretch can then be applied to each image in red, green and blue layers. This will elongate  $\mathbf{P}_k$  to  $\mathbf{P}_{ks}$  which has the same hue and saturation as  $\mathbf{P}_k$  but has increased intensity ( $OD_{ks} > OD_k$ ). The operation does not affect the hue since it only reduces the achromatism of the colour and leaves the hue information,  $\mathbf{c}$ , unchanged.

This is easy to understand if we rewrite formula (5.17) as

$$\mathbf{P}_k = \mathbf{P} - k\mathbf{a} = \mathbf{c} + \mathbf{a} - k\mathbf{a} = \mathbf{c} + (1 - k)\mathbf{a}.$$

The algebraic operations for vector formula (5.17) are

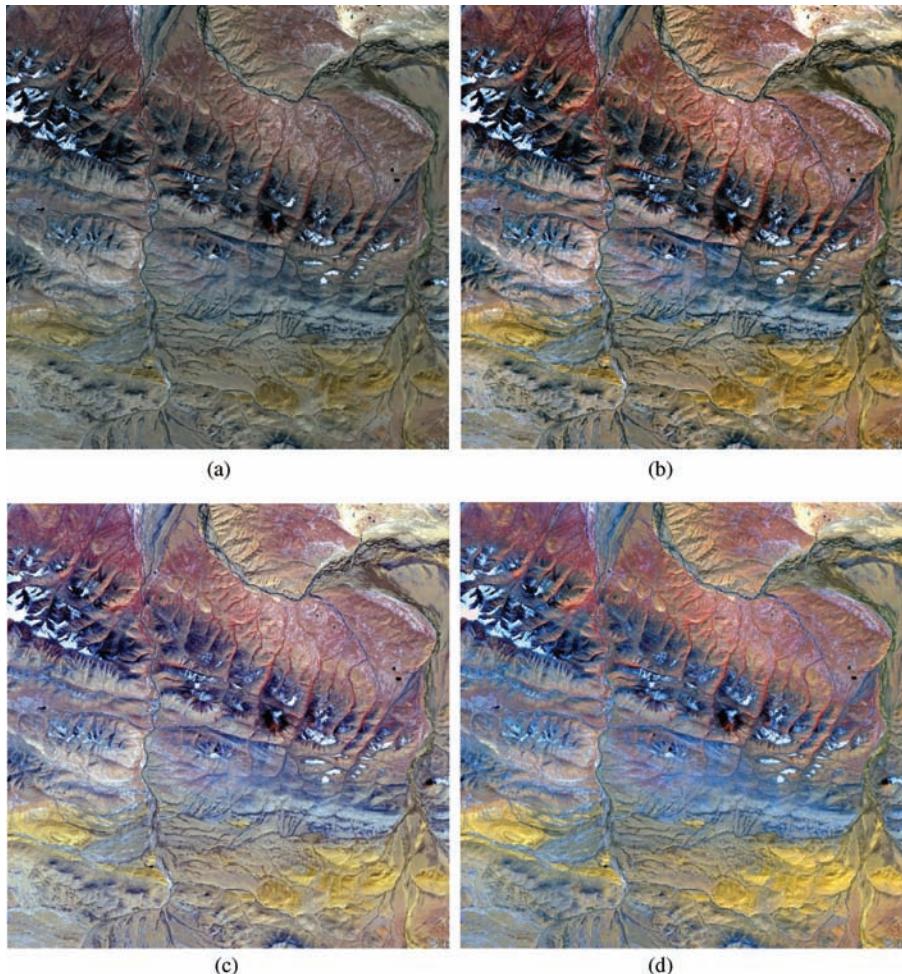
$$\begin{aligned}r_k &= r - ka = r - k \min(r, g, b) \\ g_k &= g - ka = g - k \min(r, g, b) \quad (5.18) \\ b_k &= b - ka = b - k \min(r, g, b).\end{aligned}$$

Again, as already indicated in the IHSDS, the three bands for colour composition must be well stretched (e.g. BCET or linear stretch with appropriate clipping) before the DDS is applied.

The DDS performs a decorrelation stretch essentially the same as that based on the IHS transformation as illustrated in Figure 5.4. We can prove the following properties of the DDS (refer to Section 5.6 for details):

1. DDS is controlled by the achromatic factor  $k$ .
2. For a given  $k$ , the amount of saturation stretch is dependent on the initial colour saturation; a lower saturation image is subject to stronger saturation stretch than a higher saturation image for a given  $k$ .
3. DDS does not alter the relationship between those colours with the same saturation but different intensities.
4. For colours with the same intensity but different saturation, DDS results in higher intensity for more saturated (purer) colours.

The value  $k$  is specified by users. It should be set based on the saturation level of the original colour composite. The lower the saturation of an image, the greater the  $k$  value should be given (within the range of 0–1) and  $k = 0.5$  is generally good for most cases. Figure 5.6 illustrates the initial BCET colour composite and the DDS colour composites with



**Figure 5.6** (a) A BCET standard false colour composite of Terra-1 ASTER bands 3, 2 and 1 in RGB; (b) DDS with  $k = 0.3$ ; (c)  $k = 0.5$ ; and (d)  $k = 0.7$

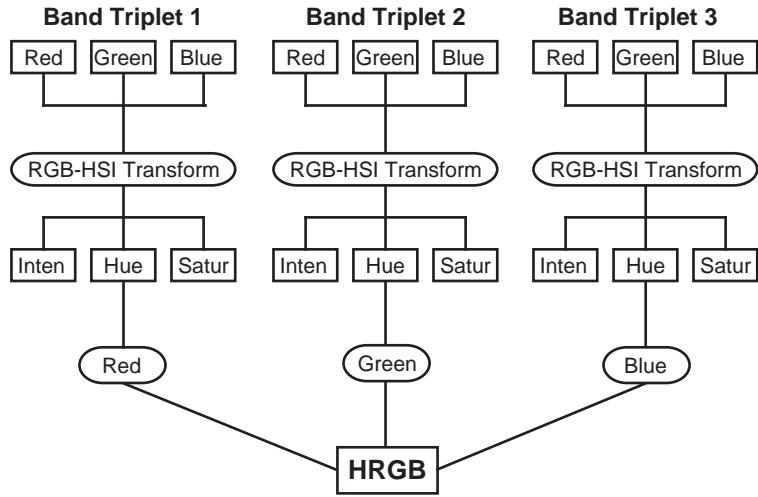
$k = 0.3, 0.5$  and  $0.7$ ; these DDS composites all show increased saturation without distortion of hues, in comparison with the original BCET colour composite, and their saturation increases with increasing  $k$ . The merits of simplicity and quantitative control of DDS are obvious.

## 5.4 Hue RGB colour composites

As shown in Figure 5.7, with the RGB–IHS transformation, three hue images can be derived from three different band triplets of a multi-spectral image. In each hue image, the brightness (the pixel DN) changes with hues which are determined by the

spectral profiles of the source bands of the triplet. If three hue images are displayed in red, green and blue using an RGB additive colour display system, a Hue RGB (HRGB) false colour composite image is produced (Liu and Moore, 1990). Colours in an HRGB image are controlled by the hue DNs of the three component hue images. An HRGB image can therefore incorporate spectral information of up to nine image bands. Pixel colours in an HRGB image are unique presentations of the spectral profiles of all the original image bands. The merits of an HRGB image are two-fold:

- It suppresses topographic shadows more effectively than ratio.



**Figure 5.7** Schematic illustration of the production of an HRGB image

where

$$\begin{aligned}
 V_t &= \sqrt{(r_t^2 + g_t^2 + b_t^2) - (r_t g_t + r_t b_t + g_t b_t)} \\
 &= \sqrt{[(nr_b)^2 + (ng_b)^2 + (nb_b)^2] - (n^2 r_b g_b + n^2 r_b b_b + n^2 g_b b_b)} \\
 &= nV_b.
 \end{aligned}$$

- It condenses and displays spectral information of up to nine image bands in a colour composite of three hue images.

From the definition of hue, it is easy to prove that  $H$  component is independent of illumination and therefore is free of topographic shadows. Suppose the irradiance upon a sunlit terrain slope,  $E_t$ , is  $n$  times that upon a terrain slope in shade,  $E_b$ . Then

$$E_t = nE_b, \quad r_t = nr_b, \quad g_t = ng_b, \quad b_t = nb_b$$

where  $r$ ,  $g$  and  $b$  represent the radiance of the three bands used for RGB colour composition.

From formula (5.2) we have

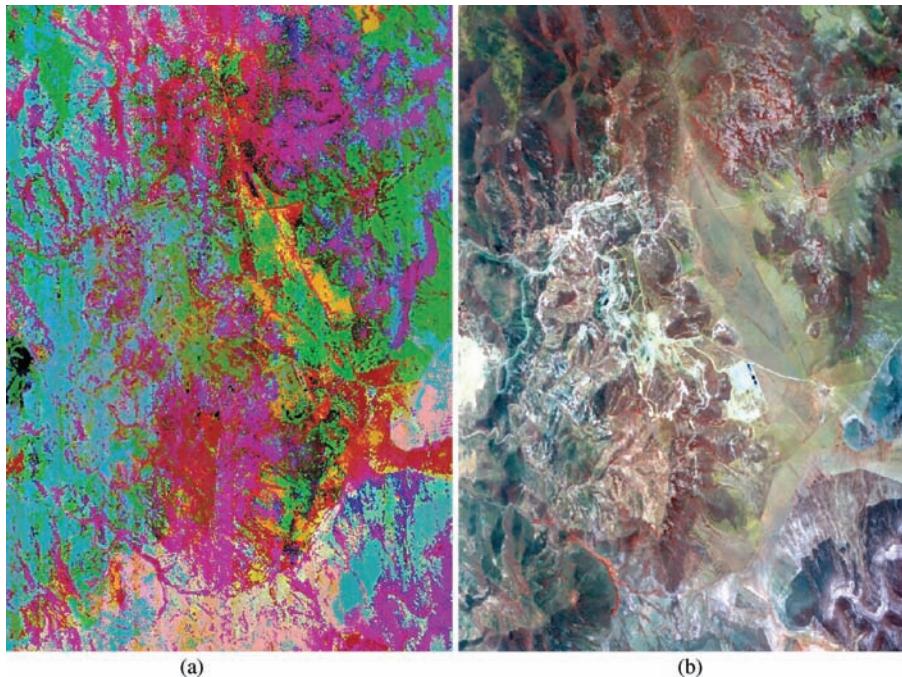
$$\begin{aligned}
 H(r_t, g_t, b_t) &= \arccos \frac{2b_t - g_t - r_t}{2V_t} \\
 &= \arccos \frac{2nb_b - ng_b - nr_b}{2nV_b} \\
 &= \arccos \frac{2b_b - g_b - r_b}{2V_b} \\
 &= H(r_b, g_b, b_b)
 \end{aligned} \tag{5.19}$$

Thus hue is independent of illumination and not affected by topographic shadows. More generally, we can prove that

$$\begin{aligned}
 H(r_i, g_i, b_i) &= H(r_j, g_j, b_j) \\
 \text{if} \\
 E_j &= nE_i + a.
 \end{aligned} \tag{5.20}$$

This equation implies that if a hue image is derived from three spectrally adjacent bands (thus the atmospheric effects on each band are not significantly different), it is little affected by shadows as well as atmospheric scattering.

With topography completely removed, an HRGB image has low SNR and it is actually difficult to interpret visually for ground objects. As shown in Figure 5.8, an HRGB image is very like a classification image without topographic features. It can therefore be used for pre-processing in preparation for classification. For visual interpretation of an HRGB image, it is advisable to use ordinary colour composites as reference images.



**Figure 5.8** (a) An HRGB colour composite of an ATM (Airborne Thematic Mapper) image: Red, hue derived from bands 10, 9, 8; Green, hue derived from bands 7, 6, 5; and Blue, hue derived from bands 4, 3, 2. (b) An ordinary colour composite of bands 8, 5 and 2 in RGB for comparison

## 5.5 \*Derivation of RGB-IHS and IHS-RGB transformations based on 3D geometry of the RGB colour cube

### 5.5.1 Derivation of RGB-IHS Transformation

As shown in Figure 5.9, the *intensity* OD is the projection of colour vector  $\mathbf{P}(r, g, b)$  onto the grey line OW or vector  $\mathbf{W}(a, a, a)$ , where  $a$  can be any value within the colour cube. Then, according to the vector projection rule,

$$I(r, g, b) = \frac{\mathbf{P} \cdot \mathbf{W}}{|\mathbf{W}|} = \frac{ra + ga + ba}{\sqrt{3a^2}} = \frac{1}{\sqrt{3}}(r + g + b). \quad (5.21)$$

The *hue* angle  $\alpha$  (or  $\angle PDE$ ) is the angle between two planes defined by triangles OBW and OPW which intercept along the grey line OW. Both planes can be defined by the three corner points of the two triangles:

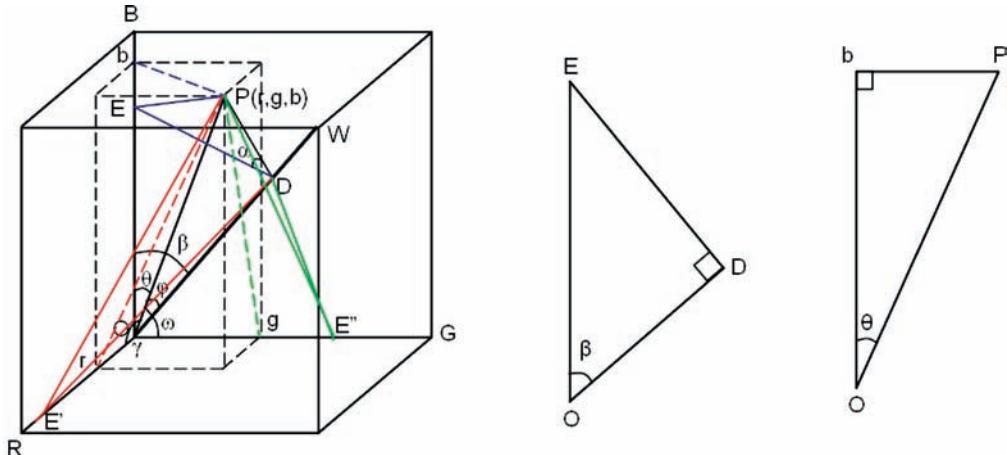
$$\text{OBW : } \begin{vmatrix} B & G & R \\ a & a & a \\ a & 0 & 0 \end{vmatrix} = 0, \quad G - R = 0 \quad (5.22)$$

$$\text{OPW : } \begin{vmatrix} B & G & R \\ a & a & a \\ b & g & r \end{vmatrix} = 0, \quad \begin{aligned} B(r-g) + G(b-r) \\ + R(g-b) = 0. \end{aligned} \quad (5.23)$$

Thus the angle between planes OBW and OPW, the hue angle  $\alpha$ , can be decided:

$$\begin{aligned} \cos \alpha &= \frac{2b - g - r}{\sqrt{2}\sqrt{(r-g)^2 + (b-r)^2 + (g-b)^2}} \\ &= \frac{2b - g - r}{2\sqrt{(r^2 + g^2 + b^2) - (rg + rb + gb)}} \end{aligned} \quad (5.24)$$

$$H(r, g, b) = \arccos \frac{2b - g - r}{2\sqrt{(r^2 + g^2 + b^2) - (rg + rb + gb)}}.$$



**Figure 5.9** The relationship between RGB and IHS in the colour cube 3D geometry

The *saturation* is the angle  $\varphi$  between colour vector  $\mathbf{P}(r, g, b)$  and grey line vector  $\mathbf{W}(a, a, a)$ . Thus, according to the vector dot product, we have

$$\begin{aligned}\cos \varphi &= \frac{\mathbf{P} \cdot \mathbf{W}}{|\mathbf{P}| |\mathbf{W}|} = \frac{a(r+g+b)}{\sqrt{r^2+g^2+b^2} \sqrt{3a^2}} \\ &= \frac{r+g+b}{\sqrt{3(r^2+g^2+b^2)}}\end{aligned}\quad (5.25)$$

$$S(r,g,b) = \arccos \frac{r+g+b}{\sqrt{3(r^2+g^2+b^2)}}.$$

### 5.5.2 Derivation of IHS–RGB transformation

Given intensity  $I$ , hue  $\alpha$  and saturation  $\varphi$ , we can derive  $R(I, \alpha, \varphi)$ ,  $G(I, \alpha, \varphi)$  and  $B(I, \alpha, \varphi)$  as depicted in Figure 5.9.

To find  $B(I, \alpha, \varphi)$ , the key is to find the angle  $\theta$  between colour vector  $\mathbf{P}$  and the B axis.

The angle between the grey line and any of the RGB axes is identical. For instance, the angle  $\beta$  between the grey line and the B axis is

$$\cos \beta = \frac{\mathbf{B} \cdot \mathbf{W}}{|\mathbf{B}| |\mathbf{W}|} = \frac{ab}{\sqrt{b^2} \sqrt{3a^2}} = \frac{1}{\sqrt{3}}. \quad (5.26)$$

As  $\cos \beta = OD/OE = 1/\sqrt{3}$ ,  $OD = I$ , thus  $OE = \sqrt{3}I$  (see the triangle OED in Figure 5.9) and

$$ED = \sqrt{OE^2 - I^2} = \sqrt{2}I.$$

The length of the colour vector  $\mathbf{P}$  is  $OP = I/\cos \varphi$  while the distance between  $\mathbf{P}$  and the grey line is  $PD = I \tan \varphi$ .

From the triangle EPD, we then have

$$\begin{aligned}EP^2 &= ED^2 + PD^2 - 2ED \times PD \cos \alpha \\ &= 2I^2 + I^2 \tan^2 \varphi - 2\sqrt{2}I^2 \tan \varphi \cos \alpha.\end{aligned}\quad (5.27)$$

From the triangle OEP, we can also find EP as

$$\begin{aligned}EP^2 &= OE^2 + OP^2 - 2OE \times OP \cos \theta \\ &= 3I^2 + \frac{I^2}{\cos^2 \varphi} - 2\sqrt{3} \frac{I^2 \cos \theta}{\cos \varphi}.\end{aligned}\quad (5.28)$$

Taking the right sides of Equations (5.27) and (5.28), we can then solve for  $\cos \theta$ :

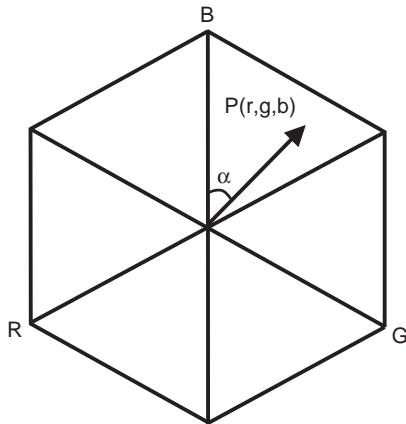
$$\cos \theta = \frac{1}{\sqrt{3}} \cos \varphi + \sqrt{\frac{2}{3}} \sin \varphi \cos \alpha. \quad (5.29)$$

As shown in the triangle ObP in Figure 5.9:

$$\begin{aligned}b &= OP \cos \theta = \frac{\cos \theta}{\cos \varphi} I \\ &= \frac{I}{\cos \varphi} \left( \frac{1}{\sqrt{3}} \cos \varphi + \sqrt{\frac{2}{3}} \sin \varphi \cos \alpha \right).\end{aligned}$$

Thus,

$$B(I, \alpha, \varphi) = b = \frac{I}{\sqrt{3}} (1 + \sqrt{2} \tan \varphi \cos \alpha). \quad (5.30)$$



**Figure 5.10** Given that hue is the angle  $\alpha$  between the B axis and colour vector  $\mathbf{P}$ , the angle between  $\mathbf{P}$  and R is  $\frac{2}{3}\pi + \alpha$  and that between  $\mathbf{P}$  and G is  $\frac{4}{3}\pi + \alpha$

If we look into the RGB colour cube along the grey line, the projection of this cube on the plane perpendicular to the grey line is a hexagon as shown in Figure 5.10. Given a hue angle  $\alpha$  of colour vector  $\mathbf{P}$  starting from the B axis, the angle between  $\mathbf{P}$  and the R axis is  $\frac{2}{3}\pi + \alpha$  and that between  $\mathbf{P}$  and G axis is  $\frac{4}{3}\pi + \alpha$ . On the other hand, the intensity  $I$  and saturation angle  $\varphi$  are both independent of the orientation of the RGB coordinate system. Thus, we can solve  $R(I, \alpha, \varphi)$  and  $G(I, \alpha, \varphi)$  in a similar way to solving  $B(I, \alpha, \varphi)$ , as above, based on the central symmetry of the grey line to the RGB axes. As shown in Figure 5.9, if we consider starting the hue angle from the R axis, then

$$OE' = OE = \sqrt{3}I \quad \text{and} \quad E'D = ED = \sqrt{2}I.$$

From triangles  $E'PD$  and  $OE'P$ , again we can establish two equations for  $E'P$ :

$$\begin{aligned} E'P^2 &= E'D^2 + PD^2 - 2E'D \times PD \cos \alpha \\ &= 2I^2 + I^2 \tan^2 \varphi - 2\sqrt{2}I^2 \tan \varphi \cos \left( \frac{2}{3}\pi + \alpha \right) \end{aligned} \quad (5.31)$$

$$\begin{aligned} E'P^2 &= OE'^2 + OP^2 - 2OE' \times OP \cos \gamma \\ &= 3I^2 + \frac{I^2}{\cos^2 \varphi} - 2\sqrt{3} \frac{I^2 \cos \gamma}{\cos \varphi} \end{aligned} \quad (5.32)$$

where  $\gamma$  is the angle between  $\mathbf{P}$  and the R axis.

Solving Equations (5.31) and (5.32) for  $\cos \gamma$ ,

$$\cos \gamma = \frac{1}{\sqrt{3}} \cos \varphi + \sqrt{\frac{2}{3}} \sin \varphi \cos \left( \frac{2}{3}\pi + \alpha \right). \quad (5.33)$$

As shown in the triangle  $OrP$  in Figure 5.9,

$$\begin{aligned} r &= OP \cos \gamma = \frac{\cos \gamma}{\cos \varphi} I \\ &= \frac{I}{\cos \varphi} \left[ \frac{1}{\sqrt{3}} \cos \varphi + \sqrt{\frac{2}{3}} \sin \varphi \cos \left( \frac{2}{3}\pi + \alpha \right) \right]. \end{aligned}$$

Finally,

$$R(I, \alpha, \varphi) = r = \frac{I}{\sqrt{3}} \left[ 1 + \sqrt{2} \tan \varphi \cos \left( \frac{2}{3}\pi + \alpha \right) \right] \quad (5.34)$$

In the same way and considering

$$\cos \left( \frac{4}{3}\pi + \alpha \right) = -\cos \left( \frac{\pi}{3} + \alpha \right),$$

we have

$$G(I, \alpha, \varphi) = g = \frac{I}{\sqrt{3}} \left[ 1 - \sqrt{2} \tan \varphi \cos \left( \frac{\pi}{3} + \alpha \right) \right]. \quad (5.35)$$

## 5.6 \*Mathematical proof of DDS and its properties

### 5.6.1 Mathematical proof of DDS

The geometrically obvious fact of the saturation stretch of the DDS in Figure 5.5 can be easily proven using simple algebra.

Let  $v = \max(r, g, b)$ . From formula (5.5), the saturation components for  $\mathbf{P}$  and  $\mathbf{P}_k$  are

$$\begin{aligned} S &= \frac{v-a}{v} = 1 - \frac{a}{v} \\ S_k &= \frac{(v-ka)-(a-ka)}{v-ka} = 1 - \frac{a-ka}{v-ka}. \end{aligned}$$

The difference between them is

$$\delta S = S_k - S = \frac{a}{v} - \frac{a-ka}{v-ka} = \frac{ka(v-a)}{v(v-ka)} \geq 0. \quad (5.36)$$

Therefore  $S_k \geq S$ .

There are three cases:

1. If  $a = 0$ , then  $S_k = S = 1$ , and colours with full saturation (pure colours) are not affected.
2. If  $a = v$ , then  $S_k = S = 0$ , and the saturation of grey tones (achromatic vectors) remains zero, though the intensity is scaled down.
3. Otherwise,  $S_k > S$  when the colour vectors between the achromatic line and the maximum saturation line are shifted (stretched) away from the grey line depending on  $k$ ,  $\mathbf{a}$  and  $\mathbf{c}$  (the effects will be further discussed in the following subsection).

### 5.6.2 The properties of DDS

DDS is independent of hue component. It enhances saturation with intensity preserved. This can be further verified by an investigation of the properties of DDS as follows.

#### 5.6.2.1 DDS is controlled by the achromatic factor $k$

The saturation increment of DDS, defined by formula (5.36), is a monotonically increasing function

of  $k$ , that is

$$\frac{d\delta S}{dk} = \frac{a(v-a)}{(v-ka)^2} > 0 \quad (5.37)$$

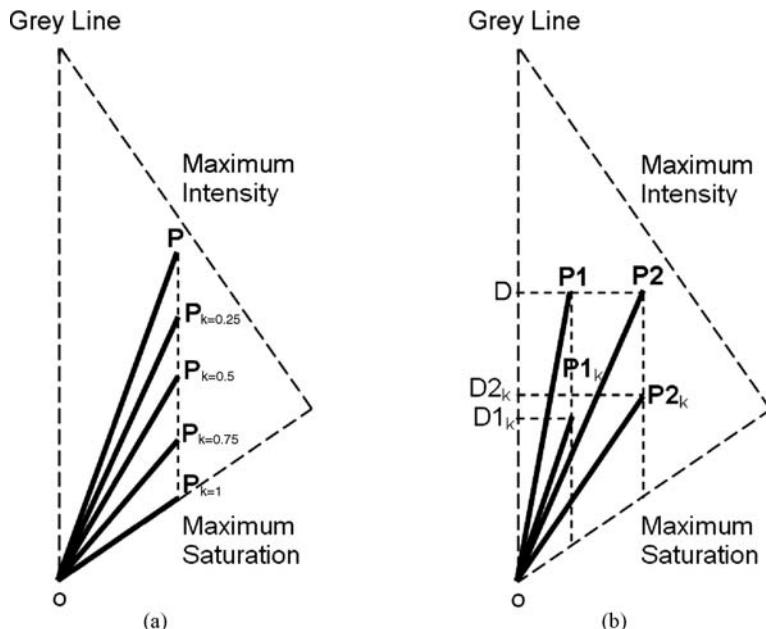
where  $v > a$  and  $0 < k < 1$ .

When  $k = 1$ ,  $\delta S$  reaches its maximum,  $\delta S_{\max} = a/v$ , and

$$S_k = S + \delta S_{\max} = \frac{v-a}{v} + \frac{a}{v} = 1.$$

This is the case for an image of chromatic component  $\mathbf{c}$ .

As shown in Figure 5.11a, a large value of  $k$  (near 1) results in a great increase in saturation ( $\delta S$ ) for a non-saturated colour vector. Such an overstretch of saturation compresses the colour vectors into a narrow range near the maximum saturation line. Conversely, a small value of  $k$  (approaching 0) has little effect on saturation. In general,  $k = 0.5$  gives an even stretch of saturation between 0 and 1. The value of  $k$  can be adjusted according to requirements. In general, a large value of  $k$  is desirable for an image with very low saturation and vice versa.



**Figure 5.11** The properties of the DDS technique: (a) variation of DDS with the achromatic factor  $k$ ; and (b) for the colours with an identical intensity, DDS results in higher intensity for a colour with higher saturation

### 5.6.2.2 For a given $k$ , the saturation stretch is dependent on colour saturation

Consider  $\delta S$  as a function of achromatic element  $a$ ; then the first derivative of  $\delta S$  for  $a$  is

$$\frac{d\delta S}{da} = \frac{1}{v} - \frac{v - kv}{(v - ka)^2}. \quad (5.38)$$

The second derivative is

$$\frac{d^2\delta S}{da^2} = -\frac{2k(1-k)v}{(v - ka)^3} < 0 \quad (5.39)$$

where  $0 < k < 1$  and  $v \geq a$ .

Therefore, as a function of  $a$ ,  $\delta S$  has a maximum when

$$\frac{d\delta S}{da} = \frac{1}{v} - \frac{v - kv}{(v - ka)^2} = 0.$$

From the above equation, we have

$$a = \frac{1 - \sqrt{1 - k}}{k} v \quad (5.40)$$

where

$$\frac{1 - \sqrt{1 - k}}{k} < 1 \quad \text{for } 0 < k < 1.$$

The saturation for the case of Equation (5.40) is then

$$S = \frac{v - a}{v} = 1 - \frac{1 - \sqrt{1 - k}}{k}. \quad (5.41)$$

Therefore, the saturation increment  $\delta S$  reaches its maximum when  $S$  satisfies Equation (5.41).

For  $k = 0.5$ , the saturation stretch reaches the maximum,  $\delta S_{\max} \approx 0.172$ , when  $S \approx 0.414$ . The saturation stretch of DDS becomes less when saturation is either greater or less than 0.414.

The criteria for the maximum stretch of saturation can be easily controlled by modifying the value of  $k$ . This property of DDS has a self-balancing effect that optimizes the stretch of the most condensed saturation range.

### 5.6.2.3 DDS does not alter the relationship between the colours with the same saturation but different intensities

Any colour with the same hue and saturation as a colour vector  $\mathbf{P}$  but different intensity can be

defined as a colour vector  $n\mathbf{P}$ :

$$n\mathbf{P} = (nr, ng, nb)$$

where  $n$  is a positive real number.

Then

$$\begin{aligned} \min(nr, ng, nb) &= na \\ n\mathbf{a} &= (na, na, na). \end{aligned}$$

In the same way as for the derivation of  $\mathbf{P}_k$ , we have

$$(n\mathbf{P})_k = n\mathbf{P} - kn\mathbf{a} = n(r - ka, g - ka, b - ka) = n\mathbf{P}_k.$$

$\mathbf{P}_k$  and  $(n\mathbf{P})_k$  have the same orientation (saturation and hue). The magnitude (intensity) relationship between them is the same as that between  $\mathbf{P}$  and  $n\mathbf{P}$ . This means that the DDS transforms one saturation value to another uniquely. DDS reduces colour intensity but does not alter the intensity relationship among colours with the same hue and saturation. For a given hue and saturation, the relative brightness of colours remains unchanged after DDS.

### 5.6.2.4 For colours with the same intensity but different saturation, DDS results in higher intensity for more saturated (purer) colours

According to the definition of colour intensity of a colour  $\mathbf{P}$  in formula (5.1), DDS shifts the colour  $\mathbf{P}$  to a colour  $\mathbf{P}_k$  with intensity

$$\begin{aligned} I_k &= \frac{1}{\sqrt{3}}(r - ka + g - ka + b - ka) \\ &= \frac{1}{\sqrt{3}}(r + g + b - 3ka) \\ &= I - \sqrt{3}ka. \end{aligned} \quad (5.42)$$

Colours with the same intensity but different saturation values have the same sum of  $r$ ,  $g$  and  $b$  but different values for  $a$  (the minimum of  $r$ ,  $g$  and  $b$ ). The higher the saturation, the smaller the value of  $a$  and the greater the value of  $I_k$  produced by the DDS according to formula (5.42). This effect is illustrated in Figure 5.11b where colours  $\mathbf{P}1$  and  $\mathbf{P}2$  have the same intensity (OD). After reducing the same proportion of their achromatic components by DDS, the intensity of the less saturated colour  $\mathbf{P}1_k$  is less than that of the more saturated colour  $\mathbf{P}2_k$  ( $OD1_k < OD2_k$ ). For information enhancement,

this property has the positive effect of increasing the variation of colours with the same intensity but different saturation in terms of differences in both intensity and saturation values.

## 5.7 Summary

The composition of three primaries, red, green and blue, produces any colour according to the tristimulus theory, while the colour quality is described as intensity, hue and saturation. The RGB–IHS transformation and the inverse transformation IHS–RGB transformation are similar to a 3D Cartesian–conical coordinate transformation and can be derived from either 3D geometry or matrix operations for coordinate rotations of the RGB colour cube.

The RGB–IHS and IHS–RGB transformations allow us to manipulate colour intensity, hue and saturation components separately and with great flexibility. One major application is the saturation-stretch-based decorrelation stretch technique that enhances image colour saturation without altering the hues of the colours. The effects are the same as reducing the inter-band correlation between the three bands for the RGB colour composition. For the same purpose, a shortcut algorithm of processing is found, the DDS (Direct Decorrelation Stretch). Based on colour vector decomposition in achromatic and chromatic components, the DDS performs the saturation stretch directly in the RGB domain without involving the RGB–IHS and IHS–RGB transformations.

Both intensity and saturation are either defined by or affected by the illumination under which

objects are imaged. The hue is, however, by definition entirely independent of illumination condition and therefore topographic shading. The hue of a colour is actually the spectral property coding. An HRGB colour composite technique is thus introduced that can code the spectral property of up to nine spectral bands into various colours to generate an information-rich colour image without the effects of topographic shadows, though the image is actually difficult for visual perception and interpretation.

## Questions

- 5.1 Use a diagram of the RGB colour cube to explain the mathematical definition and physical meaning of intensity, hue and saturation.
- 5.2 What are the value ranges of intensity, hue and saturation according to the RGB colour cube model of the RGB–IHS transformation?
- 5.3 Why is RGB–IHS a useful image processing technique?
- 5.4 Describe, with the aid of diagrams, the principle of IHS decorrelation stretch.
- 5.5 Describe the major steps of IHS decorrelation stretch.
- 5.6 What is the drawback of stretching the hue component in the IHS decorrelation stretch? Can the value range of the hue component be increased without stretching the hue component directly and, if so, how can it be achieved?
- 5.7 Use a diagram to explain the principle of DDS. In what senses are the DDS and the IHSDS similar as well as different?

# 6

## Image Fusion Techniques

The term ‘image fusion’ has become very widely used in recent years, but often to mean quite different things. Some people regard all image enhancement techniques as image fusion but in general image fusion refers specifically to techniques for integrating images or raster datasets of different spatial resolutions, or with different properties, to formulate new images. In this book, we take the latter, narrower definition and, in this chapter, following directly from the topics discussed in the previous chapters, we introduce several commonly used simple image fusion techniques for multi-resolution and multi-source image integration.

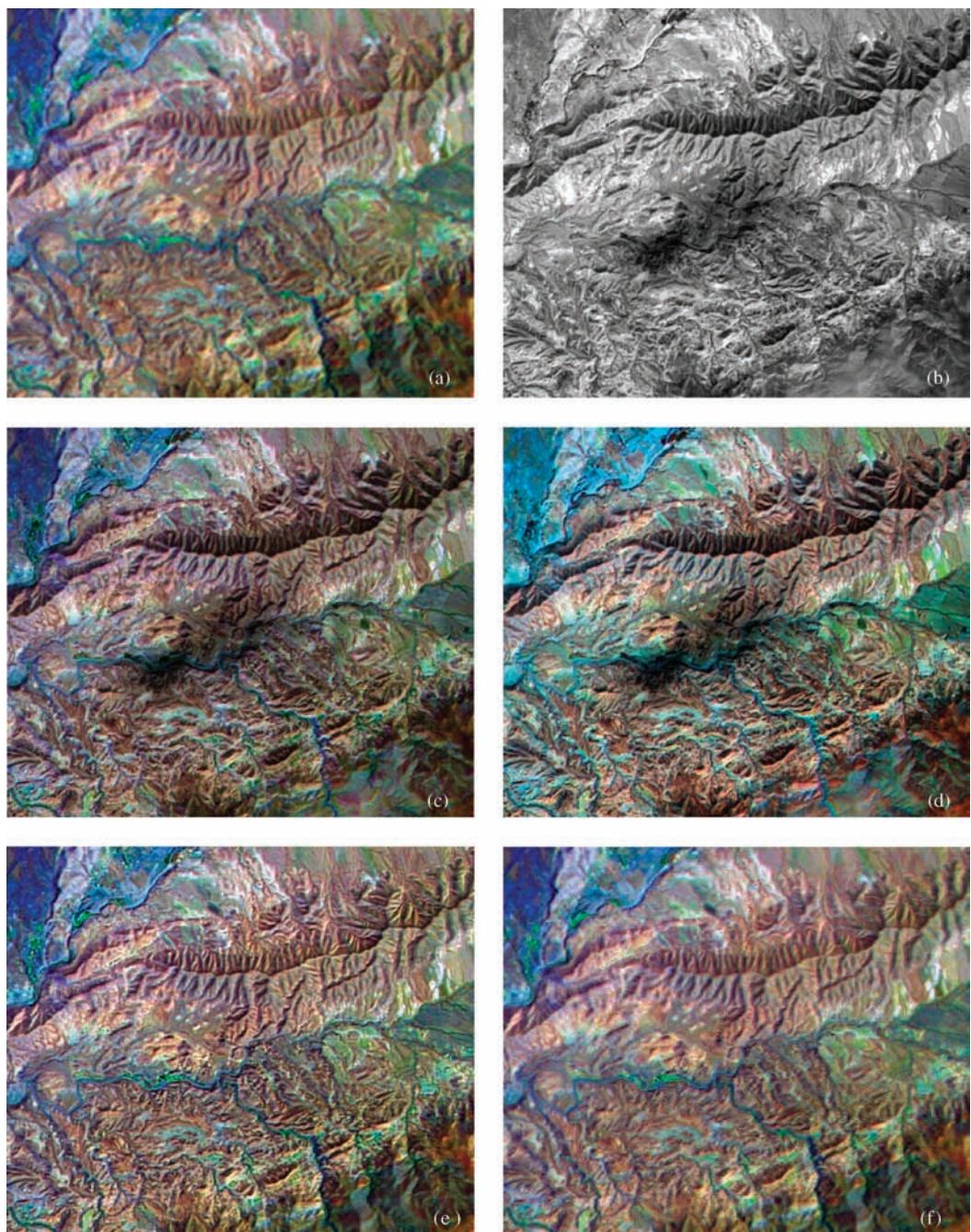
### 6.1 RGB–IHS transformation as a tool for data fusion

The RGB–IHS transformation can be used as a tool for data fusion as well as enhancement. A typical application is to fuse a low-resolution colour composite with a high-resolution panchromatic image to improve spatial resolution. With regard to optical sensor systems, image spatial resolution and spectral resolution are contradictory quantities. For a given SNR, a higher spectral resolution (narrower spectral band) is usually achieved at the cost of spatial resolution. Image fusion techniques are therefore useful for integrating a high spectral resolution image with a high spatial resolution image, such as Landsat TM (six spectral bands with

30 m resolution) and SPOT Pan (panchromatic band with 10 m resolution), to produce a fused image with high spectral and spatial resolutions. Using the RGB–IHS transformation this can be done easily by replacing the intensity component of a colour composite with a high-resolution image as follows:

1. Rectify the low-resolution colour composite image (e.g. a TM image) to the high-resolution image of the same scene (e.g. a SPOT panchromatic image). The rectification can be done the other way around, that is from high resolution to low resolution but, in some image processing software packages, the lower resolution image will need to be interpolated to the same pixel size as the high-resolution image.
2. Perform the RGB–IHS transformation on the low-resolution colour composite image.
3. Replace the intensity component,  $I$ , by the high-resolution image.
4. Perform the reverse IHS–RGB transformation.

The resultant fused image is a mixture of spectral information from the low-resolution colour composite and high spatial resolution, which better shows spatial textures and has improved overall resolution. Figure 6.1 shows a colour composite of TM bands 541 in RGB with a BCET stretch in Figure 6.1a, SPOT Pan in Figure 6.1b and the TM–SPOT Pan fusion image in Figure 6.1c. The fused image presents more detailed topographic/



**Figure 6.1** TM and SPOT Pan fusion results of several image fusion techniques: (a) a colour composite of TM bands 541RGB with BCET; (b) SPOT Pan image of the same area; (c) the IHS fusion image of TM 541 and SPOT Pan; (d) the Brovey transform fusion image; (e) the SFIM fusion image with  $5 \times 5$  smoothing filter; and (f) the SFIM fusion image with  $3 \times 3$  smoothing filter

textural information, introduced by the SPOT Pan image, and while preserving the spectral information from the three TM spectral bands; unfortunately this may occur with considerable spectral (colour) distortion. The colour distortion can be significant if the spectral range of the three TM bands for colour composition is very different from that of the panchromatic band. In this case the intensity component, calculated as the summation of the three TM bands according to formula (5.1), will be different from the SPOT Pan replacing it and, as a consequence, colour distortion is introduced.

In the same way, RGB–IHS can also be used for multi-source data integration such as the fusion of multi-spectral image data with raster geophysical or geochemical data, as outlined below:

1. Co-register the datasets to be fused.
2. RGB–IHS transformation.
3. Replacement of  $I$  component by a geophysical or geochemical dataset.
4. IHS–RGB transformation.

The resultant image contains both spectral information of the original image bands and geophysical or geochemical information as intensity variation. The interpretation of such fused images demands a thorough understanding of the input datasets. A more productive method is to use the so-called colour drape technique in which the geophysical or geochemical dataset is used as if it were a raster surface, such as a digital elevation model, with a colour composite image draped over it in a 3D perspective view. This concept will be discussed further in Part Two (Section 16.3.2).

## 6.2 Brovey transform (intensity modulation)

The Brovey transform is a shortcut to image fusion, compared with the IHS image fusion technique, and is based on direct intensity modulation. Let  $R$ ,  $G$  and  $B$  represent three image bands displayed in red, green and blue, and let  $P$  represent the image to be fused as the intensity component of the colour composite. The Brovey transform is then defined

by the following:

$$\begin{aligned} R_b &= \frac{3RP}{R+G+B} \\ G_b &= \frac{3GP}{R+G+B} \\ B_b &= \frac{3BP}{R+G+B}. \end{aligned} \quad (6.1)$$

It is obvious that the sum of the three bands in the denominator is equivalent to the intensity component of the colour composite and the Brovey transform can be simply rewritten as

$$\begin{aligned} R_b &= R \times P/I \\ G_b &= G \times P/I \\ B_b &= B \times P/I. \end{aligned} \quad (6.2)$$

The operations of the Brovey transform are therefore simply each band multiplied by the ratio of the replacement image over the intensity of the corresponding colour composite. If image  $P$  is a higher resolution image, then formulae (6.2) performs image fusion to improve spatial resolution and, if  $P$  is a raster dataset of a different source, formula (6.2) then performs multi-source data integration. The Brovey transform achieves a similar result to that of the IHS fusion technique without carrying out the whole process of RGB–IHS and IHS–RGB transformations and is thus far simpler and faster. It does, however, also introduce colour distortion, as shown in Figure 6.1d.

## 6.3 Smoothing-filter-based intensity modulation

Both the IHS and Brovey transform image fusion techniques can cause colour distortion if the spectral range of the intensity replacement (or modulation) image is different from that of the three bands in the colour composite. This problem is inevitable in colour composites that do not use consecutive spectral bands, and could become serious in vegetated and agricultural scenes if the images to be fused were acquired in different growing seasons. Preserving the original spectral properties is very important in

remote sensing applications that rely on spectral signatures, such as lithology, soil and vegetation. The spectral distortion introduced by these fusion techniques is uncontrolled and not quantified because the images for fusion are often acquired by different sensor systems, on different dates and/or in different seasons. Fusion in this context cannot therefore, in any way, be regarded as spectral enhancement and should be avoided to prevent unreliable interpretations. In seeking a spectral preservation image fusion technique that also improves spatial resolution, an image fusion technique, namely smoothing-filter-based intensity modulation (SFIM), has been developed (Liu, 2000).

### 6.3.1 The principle of SFIM

The DN value of a daytime optical image of reflective spectral band  $\lambda$  is mainly determined by two factors: the solar radiation impinging on the land surface, irradiance  $E(\lambda)$ , and the spectral reflectance of the land surface  $\rho(\lambda)$ ,  $DN(\lambda) = \rho(\lambda)E(\lambda)$ .

Let  $DN(\lambda)_{\text{low}}$  represent a DN value in a lower resolution image of spectral band  $\lambda$  and  $DN(\gamma)_{\text{high}}$  the DN value of the corresponding pixel in a higher resolution image of spectral band  $\gamma$ , and assume that the two images are taken in similar solar illumination conditions (such as the case of TM and SPOT). Then

$$DN(\lambda)_{\text{low}} = \rho(\lambda)_{\text{low}}E(\lambda)_{\text{low}} \quad \text{and}$$

$$DN(\gamma)_{\text{high}} = \rho(\gamma)_{\text{high}}E(\gamma)_{\text{high}}.$$

After co-registering the lower resolution image precisely to the higher resolution image and meanwhile interpolating the lower resolution image to the same pixel size of the higher resolution image, the SFIM technique is defined as

$$\begin{aligned} DN(\lambda)_{\text{sim}} &= \frac{DN(\lambda)_{\text{low}} \times DN(\gamma)_{\text{high}}}{DN(\gamma)_{\text{mean}}} \\ &= \frac{\rho(\lambda)_{\text{low}}E(\lambda)_{\text{low}} \times \rho(\gamma)_{\text{high}}E(\gamma)_{\text{high}}}{\rho(\gamma)_{\text{low}}E(\gamma)_{\text{low}}} \\ &\approx \rho(\lambda)_{\text{low}}E(\lambda)_{\text{high}} \end{aligned} \quad (6.3)$$

where  $DN(\lambda)_{\text{sim}}$  is the simulated higher resolution pixel corresponding to  $DN(\lambda)_{\text{low}}$  and  $DN(\gamma)_{\text{mean}}$  the local mean of  $DN(\gamma)_{\text{high}}$  over a neighbourhood equivalent to the resolution of  $DN(\lambda)_{\text{low}}$ .

For a given solar radiation, irradiance upon a land surface is controlled by topography. If the two images are quantified to the same DN range, we can presume that  $E(\lambda) \approx E(\gamma)$  for a given resolution because both vary with topography in the same way as denoted in Equation 3.27. We can also presume that  $\rho(\gamma)_{\text{low}} \approx \rho(\gamma)_{\text{high}}$  if there is no significant spectral variation within the neighbourhood for calculating  $DN(\gamma)_{\text{mean}}$ . Thus, in

$$\frac{\rho(\lambda)_{\text{low}}E(\lambda)_{\text{low}} \times \rho(\gamma)_{\text{high}}E(\gamma)_{\text{high}}}{\rho(\gamma)_{\text{low}}E(\gamma)_{\text{low}}},$$

$E(\lambda)_{\text{low}}$  and  $E(\gamma)_{\text{low}}$  cancel each other;  $\rho(\gamma)_{\text{low}}$  and  $\rho(\gamma)_{\text{high}}$  also cancel each other; and  $E(\gamma)_{\text{high}}$  can be replaced by  $E(\lambda)_{\text{high}}$ . We then have the final simple solution of formula (6.3).

The local mean  $DN(\gamma)_{\text{mean}}$  is calculated for every pixel of the higher resolution image using a convolution smoothing filter. The filter kernel size is decided by the resolution ratio between the higher and lower resolution images. For instance, to fuse a 30 m resolution TM band image with a 10 m resolution SPOT Pan image, the minimum smoothing filter kernel size for calculating the local mean of the SPOT Pan image pixels is  $3 \times 3$  defined as

$$\frac{1}{9} \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}.$$

The image of  $DN(\gamma)_{\text{mean}}$  is equivalent to the image of  $DN(\lambda)_{\text{low}}$  in topography and texture because they both have a pixel size of the higher resolution image and a spatial resolution of the lower resolution image. The crucial approximation,  $E(\lambda) \approx E(\gamma)$ , for simplifying formula (6.3), therefore stands and the approach of SFIM is valid. The final result, the image of  $DN(\lambda)_{\text{sim}}$ , is a product of the higher resolution topography and texture,  $E(\lambda)_{\text{high}}$ , introduced from the higher

resolution image, and the lower resolution spectral reflectance of the original lower resolution image,  $\rho(\lambda)_{\text{low}}$ . It is therefore independent of the spectral property of the higher resolution image used for intensity modulation. In other words, SFIM is a spectral preservation fusion technique. This is the major advantage of SFIM over the IHS and Brovey transform fusion techniques.

Since the spectral difference between the lower and the higher resolution images is not fundamental to the operations, formula (6.3) can be more concisely presented as a general SFIM processing algorithm:

$$\text{IMAGE}_{\text{SFIM}} = \frac{\text{IMAGE}_{\text{low}} \times \text{IMAGE}_{\text{high}}}{\text{IMAGE}_{\text{mean}}} \quad (6.4)$$

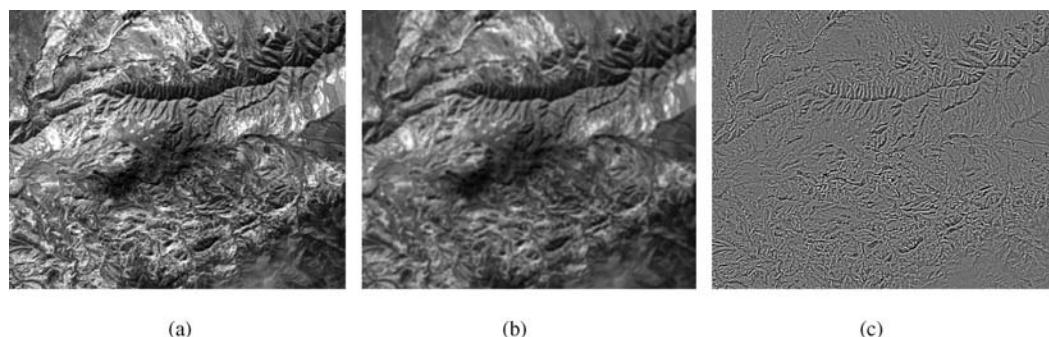
where  $\text{IMAGE}_{\text{low}}$  is a pixel of a lower resolution image co-registered to a higher resolution image of  $\text{IMAGE}_{\text{high}}$ , and  $\text{IMAGE}_{\text{mean}}$  a smoothed pixel of  $\text{IMAGE}_{\text{high}}$  using an averaging filter over a neighbourhood equivalent to the actual resolution of  $\text{IMAGE}_{\text{low}}$ .

The ratio between  $\text{IMAGE}_{\text{high}}$  and  $\text{IMAGE}_{\text{mean}}$  in formula (6.4) cancels the spectral and topographical contrast of the higher resolution image and retains the higher resolution edges only, as illustrated by a SPOT Pan image in Figure 6.2. SFIM can thus be understood as a lower resolution image directly modulated by higher resolution edges and the result is independent of the contrast and spectral variation of the higher resolution image. SFIM is therefore reliable for the spectral properties as well as for the contrast of the original lower resolution image.

### 6.3.2 Merits and limitation of SFIM

Figure 6.1 illustrates the TM–SPOT Pan fusion results produced by the IHS, Brovey transform and SFIM fusion techniques. It is clear that the SFIM result (Figure 6.1e) demonstrates the highest spectral fidelity to the original TM band 541 colour composite (Figure 6.1a) showing no noticeable colour differences, while the fusion results of both the IHS and Brovey transforms (Figure 6.1c and d respectively) present considerable colour distortion as well as contrast changes. In particular, a patch of thin cloud and shadow at the bottom right and central part of the SPOT Pan image are fused into the IHS and Brovey transform fusion images, which are not shown in the SFIM fusion images because they are cancelled out by the ratio between the original SPOT Pan image and its smoothed version as shown in Figure 6.2c.

SFIM is sensitive to the accuracy of image co-registration. Edges with imperfect co-registration will become slightly blurred because the cancellation of  $E(\lambda)_{\text{low}}$  and  $E(\gamma)_{\text{low}}$  in formula (6.3) is no longer perfect in such a case. This problem can be eased by using a smoothing filter with a larger kernel than the resolution ratio. In such a case,  $E(\gamma)_{\text{low}}$  represents lower frequency information than  $E(\lambda)_{\text{low}}$  in formula (6.3). The division between the two does not lead to a complete cancellation and the residual is the high-frequency information of the lower resolution image (relating to edges). Thus, in the fused image, the main edges appearing in both images will be sharpened while the subtle textural patterns,



**Figure 6.2** (a) Original SPOT Pan image; (b) smoothed SPOT Pan image with a  $5 \times 5$  smoothing filter; and (c) the ratio image between (a) and (b)

which are recognizable only in the higher resolution image, will be retained. Figure 6.1e is processed using SFIM with a  $5 \times 5$  smoothing filter; the blurring effects are effectively suppressed while the spatial details are significantly improved. In comparison, SFIM with a  $3 \times 3$  smoothing filter in Figure 6.1f is rather blurred, which cancels out the improvement of spatial resolution. Thus a filter kernel at least one step larger than the resolution ratio between the low- and high-resolution images is recommended. Another way to improve the SFIM fusion quality is to achieve precise pixel-to-pixel image co-registration, and this will be introduced in Chapter 9.

Attention should be paid to the following issues:

1. As already mentioned, it is important that for the SFIM operations defined by Equation 6.4, the lower resolution image must be interpolated to the same pixel size as the higher resolution image by the co-registration process; that is,  $\text{IMAGE}_{\text{low}}$  must have the same pixel size as  $\text{IMAGE}_{\text{high}}$  even though it is in a lower resolution. For the lower resolution image interpolation, simple pixel duplication must be avoided; instead, bilinear, biquadratic or bicubic resampling should be applied.
2. As SFIM is based on a solar radiation model, the technique is not applicable to the fusion of images with different illumination and imaging geometry, such as TM and ERS-1 SAR, or to integrate multi-source raster datasets.
3. If the spectral range of the higher resolution image is the same as that of the lower resolution colour composite and they are taken in similar solar radiation conditions, none of the three fusion techniques will introduce significant spectral (colour) distortion. In this case, the IHS and Brovey transform fusion techniques are preferable to SFIM for producing sharper images.

## 6.4 Summary

Image fusion is a very active research field. In this chapter, we introduced the three simplest and most

popular image fusion techniques that can be performed by commonly used image processing software packages.

For image fusion aiming to improve the spatial resolution of multi-spectral images using panchromatic images, the minimization of spectral distortion is one of the most important issues, while achieving the sharpest image textures is another, not to mention economizing on the processing speed. SFIM provides a spectral preservation image fusion solution but its high requirements in terms of image co-registration accuracy often result in slightly blurred edge textures. This weakness can be easily amended by using a larger smoothing filter. A more thorough solution to overcome the weakness is a new method developed for pixel-to-pixel image co-registration at sub-pixel accuracy as introduced in Chapter 9. On the other hand, the RGB–IHS and Brovey transform-based techniques remain popular because of their simplicity and robustness, though spectral distortion is often unavoidable.

In seeking robust spectral preservation image fusion techniques, some techniques have been developed based on wavelet transforms. Considering the current general availability, robustness, mathematic complexity and processing efficiency of such wavelet-transform-based techniques, we decide not to cover this branch so as to keep the contents concise and essential.

## Questions

- 6.1 How could you improve the spatial resolution of a 30 m resolution TM colour composite with a 10 m resolution SPOT panchromatic image, using RGB–IHS transformation and the Brovey transform?
- 6.2 Explain the major problem of image fusion using the RGB–IHS and Brovey transformations.
- 6.3 Describe the principle and derivation of the SFIM method, explaining why SFIM is a spectral preservation image fusion method.
- 6.4 What is the main problem for SFIM and how should it be dealt with?

# 7

## Principal Component Analysis

Principal component analysis (PCA) is a general method of analysis for correlated multi-variable datasets. Remotely sensed multi-spectral imagery is typical of such datasets for which PCA is an effective technique for spectral enhancement and information manipulation. PCA is based on linear algebraic matrix operations and multi-variable statistics. Here we focus on the principles of the PCA technique and its applications and avoid going into the mathematical details since these comprise fairly standard linear algebraic algorithms which are implemented in most image processing software packages.

Relying on the concept of PCA as a coordinate rotation, we expand our discussion to the general concept of physical-property-orientated image coordinate transformation. This discussion also leads to the widely used tasseled cap transformation in the derivation of multi-spectral indices of brightness, greenness and wetness.

PCA can effectively concentrate the maximum information of many correlated image spectral bands into a few uncorrelated principal components and therefore can reduce the size of a dataset and enable effective image RGB display of its information. This links to the statistical methods for band selection that aim at selecting optimum band triplets with minimal inter-band correlation and maximum information content.

### 7.1 Principle of PCA

As shown in Table 7.1, the six reflective spectral bands of a TM image are highly correlated. For instance, the correlation between band 5 and band 7 is 0.993. This means that there is 99.3% information redundancy between these two bands and only 0.7% of unique information! This is the general case for multi-spectral Earth observation image data because topography represents the image features common to all bands. The narrower the spectral ranges of the image bands, the higher the correlation between the adjacent bands. As such, multi-spectral imagery is not efficient for information storage.

Consider an  $m$ -band multi-spectral image as an  $m$ -dimensional raster dataset in an  $m$  dimension orthogonal coordinate system, forming an  $m$ -dimensional ellipsoid cluster. Then the coordinate system is oblique to the axes of the ellipsoid data cluster if the image bands are correlated. The axes of the data ellipsoid cluster formulate an orthogonal coordinate system in which the same image data are represented by  $n$  ( $n \leq m$ ) independent principal components. In other words, the *principal components* (PCs) are the image data representation in the coordinate system formed by the axes of the ellipsoid data cluster. Thus, PCA is a coordinate rotation operation to rotate the coordinate system of the original image bands to match the axes of the

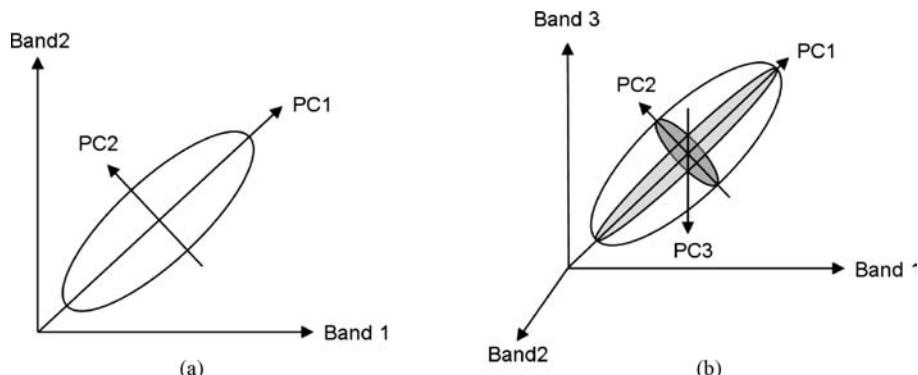
**Table 7.1** Correlation matrix of bands 1–5 and 7 of a TM sub-scene

Correlation	<i>TM1</i>	<i>TM2</i>	<i>TM3</i>	<i>TM4</i>	<i>TM5</i>	<i>TM7</i>
<i>TM1</i>	1.000	0.962	0.936	0.881	0.839	0.850
<i>TM2</i>	0.962	1.000	0.991	0.965	0.933	0.941
<i>TM3</i>	0.936	0.991	1.000	0.979	0.955	0.964
<i>TM4</i>	0.881	0.965	0.979	1.000	0.980	0.979
<i>TM5</i>	0.839	0.933	0.955	0.980	1.000	0.993
<i>TM7</i>	0.850	0.941	0.964	0.979	0.993	1.000

ellipsoid of the image data cluster. As shown by the 2D illustration in Figure 7.1a, suppose the image data points form an elliptic cluster; the aim of PCA is to rotate the orthogonal coordinate system of band 1 and band 2 to match the two axes of the ellipsoid, PC1 and PC2. The coordinates of each data point in the PC coordinate system will be the DNs of the corresponding pixels in the PC images. The first PC is represented by the longest axis of the data cluster, the second PC by the second longest, and so on. The axes representing high-order PCs may be too short to represent any substantial information and then the apparent  $m$ -dimensional ellipsoid is effectively degraded to  $n$  ( $n < m$ ) independent dimensions. For instance, as shown in Figure 7.1b, the 3D data cluster is effectively 2D as the PC3 axis is very short, representing little independent information. The same data can then be effectively represented by PC1 and PC2 in a 2D coordinate system with little information loss. In this way, PCA reduces

image dimensionality and represents nearly the same image information with fewer independent dimensions in a smaller dataset without redundancy. In summary, PCA is a linear transformation converting  $m$  correlated dimensions to  $n$  ( $n \leq m$ ) independent (uncorrelated) dimensions. This is equivalent to a coordinate rotation transform to rotate the original  $m$  axes oblique to the ellipsoid data cluster to match the orientation of the axes of the ellipsoid in  $n$  independent dimensions, and thus the image data represented by each dimension are orthogonal to (independent of) all the other dimensions. For image processing, PCA generates uncorrelated PC images from the originally correlated image bands.

Let  $\mathbf{X}$  represent an  $m$ -band multi-spectral image; then its covariance matrix  $\Sigma_x$  is a full representation of the  $m$ -dimensional ellipsoid cluster of the image data. The covariance matrix is a non-negative definite matrix and it is symmetric along its major diagonal. Such a matrix can be converted into a



**Figure 7.1** Principle of principal component transformation: (a) a 2D case of PCA; and (b) a 3D case of PCA. The 3D cluster is effectively two dimensional as the value range of PC3 is very narrow and the data distribution is mainly in an elliptic plate

diagonal matrix via basic matrix operations. The elements on the major diagonal of the covariance matrix are the variance of each image band, while the symmetrical elements off the major diagonal are the covariance between two different bands. For instance, the covariance matrix of a four-band image is

$$\begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} & \sigma_{24} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} & \sigma_{34} \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_{44} \end{pmatrix}.$$

For the elements not on the major diagonal in this matrix,  $\sigma_{ij} = \sigma_{ji}$ . The elements  $\sigma_{12}$  and  $\sigma_{21}$  both are the covariance between band 1 and band 2 and so on. If band 1 and band 2 are independent, then their covariance  $\sigma_{12} = \sigma_{21} = 0$ . This means that independent variables in a multi-dimensional space should have a diagonal covariance matrix. Thus an image dataset of  $n$  independent PCs should have a diagonal covariance matrix.

In mathematics, PCA is simply to find a transformation  $\mathbf{G}$  that diagonalizes the covariance matrix  $\Sigma_x$  of the  $m$ -band image  $\mathbf{X}$  to produce an  $n$  PC image  $\mathbf{Y}$  with a diagonal covariance matrix  $\Sigma_y$ . The rank of  $\Sigma_y$  is  $n$  and  $n = m$  if the  $m$  bands are independent, and  $\Sigma_x$  is then a full-rank matrix, otherwise  $n < m$  with reduced dimensionality.

The covariance matrix of an  $m$ -band image  $\mathbf{X}$  is defined as

$$\begin{aligned} \Sigma_x &= \mathbb{E}\left\{(\mathbf{x} - \mathbf{m}_x)(\mathbf{x} - \mathbf{m}_x)^T\right\} \\ &\approx \frac{1}{N-1} \sum_{j=1}^N (\mathbf{x}_j - \mathbf{m}_x)(\mathbf{x}_j - \mathbf{m}_x)^T \quad (7.1) \end{aligned}$$

where  $\mathbf{x}_j = (x_{j1}, x_{j2}, \dots, x_{jm})^T$  ( $\mathbf{x}_j \in \mathbf{x}, j = 1, 2, \dots, N$ ) is any  $m$ -dimensional pixel vector of an  $m$ -band image  $\mathbf{X}$ ,  $N$  the total number of pixels in the image  $\mathbf{X}$  and  $\mathbf{m}_x$  the mean vector of the image  $\mathbf{X}$ . The operation  $\mathbb{E}$  is a mathematical expectation.

$$\mathbf{m}_x = \mathbb{E}\{\mathbf{x}\} = \frac{1}{N-1} \sum_{j=1}^N \mathbf{x}_j. \quad (7.2)$$

Since the covariance matrix  $\Sigma_x$  is a symmetrical, non-negative definite matrix, there exists a linear

transformation  $\mathbf{G}$  that diagonalizes  $\Sigma_x$ . Let

$$\mathbf{y} = \mathbf{G}\mathbf{x} \quad (7.3)$$

subject to the constraint that the covariance matrix of  $\mathbf{y}_j \in \mathbf{y}, j = 1, 2, \dots, N$  is diagonal. In  $\mathbf{Y}$  space the covariance matrix is, by definition,

$$\Sigma_y = \mathbb{E}\{(\mathbf{y} - \mathbf{m}_y)(\mathbf{y} - \mathbf{m}_y)^T\} \quad (7.4)$$

where  $\mathbf{m}_y$  is the mean vector of the transformed image  $\mathbf{Y}$ . Thus we have

$$\mathbf{m}_y = \mathbb{E}\{\mathbf{y}\} = \mathbb{E}\{\mathbf{G}\mathbf{x}\} = \mathbf{G}\mathbb{E}\{\mathbf{x}\} = \mathbf{G}\mathbf{m}_x \quad (7.5)$$

$$\begin{aligned} \Sigma_y &= \mathbb{E}\{(\mathbf{G}\mathbf{x} - \mathbf{G}\mathbf{m}_x)(\mathbf{G}\mathbf{x} - \mathbf{G}\mathbf{m}_x)^T\} \\ &= \mathbf{G}\mathbb{E}\{(\mathbf{x} - \mathbf{m}_x)(\mathbf{x} - \mathbf{m}_x)^T\}\mathbf{G}^T \\ &= \mathbf{G}\Sigma_x\mathbf{G}^T. \end{aligned} \quad (7.6)$$

As  $\Sigma_y$  is the diagonal matrix derived from  $\Sigma_x$ , according to the rules of matrix operations we can prove that the transformation  $\mathbf{G}$  is the  $n \times m$  transposed matrix of the eigenvectors of  $\Sigma_x$ :

$$\mathbf{G} = \begin{pmatrix} g_{11} & g_{12} & \cdots & g_{1m} \\ g_{21} & g_{22} & \cdots & g_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ g_{n1} & g_{n2} & \cdots & g_{nm} \end{pmatrix} = \begin{pmatrix} \mathbf{g}_1^T \\ \mathbf{g}_2^T \\ \vdots \\ \mathbf{g}_n^T \end{pmatrix}. \quad (7.7)$$

$\Sigma_y$  is a diagonal matrix with the eigenvalues of  $\Sigma_x$  as non-zero elements along the diagonal:

$$\Sigma_y = \begin{pmatrix} \lambda_1 & & & 0 \\ & \lambda_2 & & \\ & & \ddots & \\ 0 & & & \lambda_n \end{pmatrix} \quad (7.8)$$

$\lambda_1 > \lambda_2 > \dots > \lambda_n.$

The eigenvalue  $\lambda_i$  is the variance of the PC <sub>$i$</sub>  image and is proportional to the information contained in PC <sub>$i$</sub> . As indicated in (7.8), the information content decreases with the increment in the PC rank.

In computing, the key operation of PCA is to find eigenvalues of  $\Sigma_x$  from which the eigenvector matrix  $\mathbf{G}$  is derived. The eigenvalues of  $\Sigma_x$  can be calculated from its *characteristic equation*:

$$|\Sigma_x - \lambda\mathbf{I}| = 0 \quad (7.9)$$

where  $\mathbf{I}$  is an  $m$ -dimension identity matrix.

**Table 7.2** The covariance matrix of bands 1–5 and 7 of a TM sub-scene

Covariance	<i>TM1</i>	<i>TM2</i>	<i>TM3</i>	<i>TM4</i>	<i>TM5</i>	<i>TM7</i>
<i>TM1</i>	232.202	196.203	305.763	348.550	677.117	345.508
<i>TM2</i>	196.203	178.980	284.415	335.185	660.570	335.997
<i>TM3</i>	305.763	284.415	460.022	545.336	1083.993	551.367
<i>TM4</i>	348.550	335.185	545.336	674.455	1347.927	678.275
<i>TM5</i>	677.117	660.570	1083.993	1347.927	2802.914	1402.409
<i>TM7</i>	345.508	335.997	551.367	678.275	1402.409	711.647

An eigenvector of matrix  $\Sigma_x$  is defined as a vector  $\mathbf{g} (\mathbf{g} \in \mathbf{G})$  that satisfies

$$\Sigma_x \mathbf{g} = \lambda \mathbf{g} \quad \text{or} \quad (\Sigma_x - \lambda \mathbf{I}) \mathbf{g} = 0. \quad (7.10)$$

This formula is called the *characteristic polynomial* of  $\Sigma_x$ . Thus, once the *i*th eigenvalue  $\lambda_i$  is known, then the *i*th eigenvector  $\mathbf{g}_i$  is determined. There are several standard computing algorithms for numerical solutions of the characteristic equation (7.9) but the mathematics is beyond the scope of this book.

Eigenvector  $\mathbf{G}$  determines how each PC is composed from the original image bands. In fact, each PC image is a linear combination (a weighted summation) of the original image bands:

$$\text{PC}_i = \mathbf{g}_i^T \mathbf{X} = \sum_{k=1}^m g_{ik} \text{Band}_k \quad (7.11)$$

where  $g_{ik}$  is the element of  $\mathbf{G}$  at the *i*th row and *k*th column, or the *k*th element of the *i*th eigenvector  $\mathbf{g}_i^T = (g_{i1}, g_{i2}, \dots, g_{ik}, \dots, g_{im})$ .

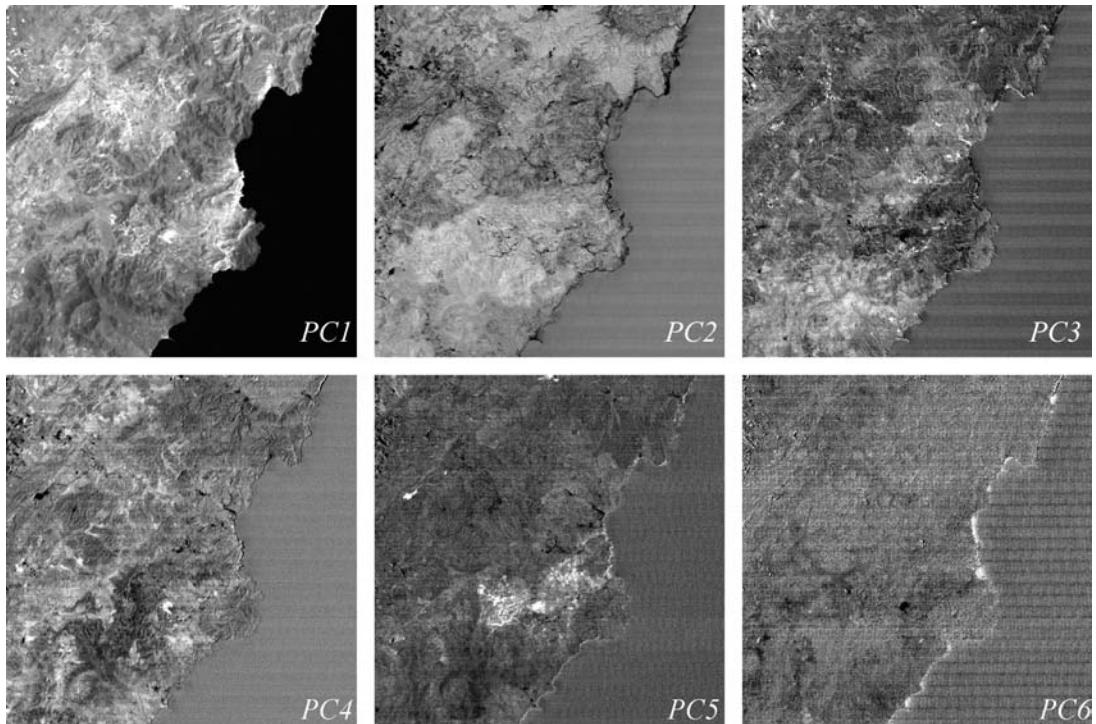
## 7.2 Principal component images and colour composition

PC images are useful for reducing data dimensionality, condensing topographic and spectral information, improving image colour presentation and enhancing specific spectral features. Here we discuss some characteristic of PC images using an example. The covariance matrix and eigenvector matrix of six reflective spectral bands of a small sub-scene of a Landsat TM image are presented in Tables 7.2 and 7.3 and the inter-band correlation matrix is shown in Table 7.1. The six PC images are shown in Figure 7.2. We make the following observations:

1. The elements of  $\mathbf{g}_1$  are all positive and therefore PC1 (Figure 7.2a) is a weighted average of all the original image bands. In this sense, it resembles a broad spectral range panchromatic image. It has a very large eigenvalue 4928.731 (PC1 variance) and accounts for 97.4% of the information from

**Table 7.3** The eigenvector matrix and eigenvalues of the covariance matrix of bands 1–5 and 7 of a TM sub-scene

Eigenvectors	PC1	PC2	PC3	PC4	PC5	PC6
<i>TM1</i>	0.190	-0.688	-0.515	-0.260	-0.320	-0.233
<i>TM2</i>	0.183	-0.362	0.032	0.050	0.136	0.902
<i>TM3</i>	0.298	-0.418	0.237	0.385	0.638	-0.354
<i>TM4</i>	0.366	-0.136	0.762	-0.330	-0.389	-0.079
<i>TM5</i>	0.751	0.433	-0.296	-0.318	0.242	0.013
<i>TM7</i>	0.378	0.122	-0.093	0.756	-0.511	0.011
Eigenvalues	4928.731	102.312	15.581	9.011	3.573	1.012
Information	97.4%	2.02%	0.31%	0.18%	0.07%	0.02%



**Figure 7.2** PC images derived from six reflective spectral bands of a sub-scene of a TM image. The PC1–PC6 images are arranged from top left to bottom right

- all six bands. For a fixed DN range, more information means a higher SNR. This conforms to the conclusion that image summation increases SNR as stated in Section 3.1.
2. PC1 concentrates features common to all six bands. For Earth observation satellite images, this common information is usually topography.
  3. The elements of  $\mathbf{g}_i$  ( $i > 1$ ) are usually a mixture of positive and negative values and thus PC images of higher rank ( $>1$ ) are linear combinations of positive and negative images of the original bands.
  4. The higher ranked PCs lack topographic features and show more spectral variation. They all have significantly smaller eigenvalues (PC variances) than PC1. The eigenvalues decrease rapidly with the increment of PC rank and so have progressively lower SNRs, which is illustrated by their increasingly noisy appearance. The PC6 image is almost entirely noise and contains little information, as indicated by the very small variance

1.012. In this sense, PC6 can be disregarded from the dataset and thus the effective dimensionality is reduced to five from the original six with negligible information loss of 0.02%.

We can look at individual PC images or display three PCs as a colour composite. As PC1 is mainly topography, colour composites excluding PC1 may better present spectral information with topography subdued. PCs represent condensed and independent image information and therefore produce more colourful (i.e. informative) colour composites. However, here we have a problem in that a PC is a linear combination of the original spectral bands, so its relationship to the original spectral signatures of targets representing ground objects is no longer apparent.

To solve this problem, a feature-oriented PC selection (FPCS) method for colour composition was proposed by Crosta and Moore (1989). The technique provides a simple way to select PCs from

the spectral signatures of significant spectral targets (e.g. minerals) to enhance the spectral information of these minerals in the PC colour composite. The technique involves examination of the eigenvectors to identify the contributions from original bands (either negative or positive) to each PC. Specific PCs can then be selected on the basis of the major contributors that are likely to display the desired targets (spectral features).

Let us look at the eigenvectors in Table 7.3. PC3 is dominated by large positive loading (0.762) from *TM4* caused by the high reflectance of vegetation in NIR and large negative loading (-0.515) of the blue-band *TM1*. The red-band *TM3* and the SWIR band *TM5* give the second largest positive and negative contribution respectively. PC3 therefore highlights vegetation and particularly vegetation on red soils. The largest positive contribution to PC4 is from the clay absorption SWIR band *TM7* (0.756) and the other elements of PC4 include a positive contribution from *TM3* and negative contributions of similar amounts from *TM1*, *TM4* and *TM5*. PC4 therefore enhances ferric iron and lacks the signature of clay minerals. PC5 can be considered as a combination of the difference between *TM5* and *TM7*, highlighting clay minerals, and between *TM3* and *TM1*, highlighting iron oxides, together with a negative contribution from vegetation. PC5 is therefore effective for indicating hydrothermal alteration minerals by its strong co-occurrence of clay minerals and iron oxide. Figure 7.3a is a colour composite displaying PC4 in red, PC3 in green and PC5 in blue. Apart from showing vegetation in green and red soils/regoliths in red, the image effectively highlights hydrothermal alteration zones of a known epithermal gold deposit distinctively in blue. Geologically the alteration zone is characterized by high concentrations of alteration-induced clay minerals and gossaniferous iron oxides on the surface of outcrops.

### 7.3 Selective PCA for PC colour composition

The eigenvalues representing the variances of PCs in Table 7.3 indicate that for commonly used, remotely sensed multi-spectral image data, a very large portion of information (data variance) is concentrated in PC1 and that this relates to the irradiance variation on

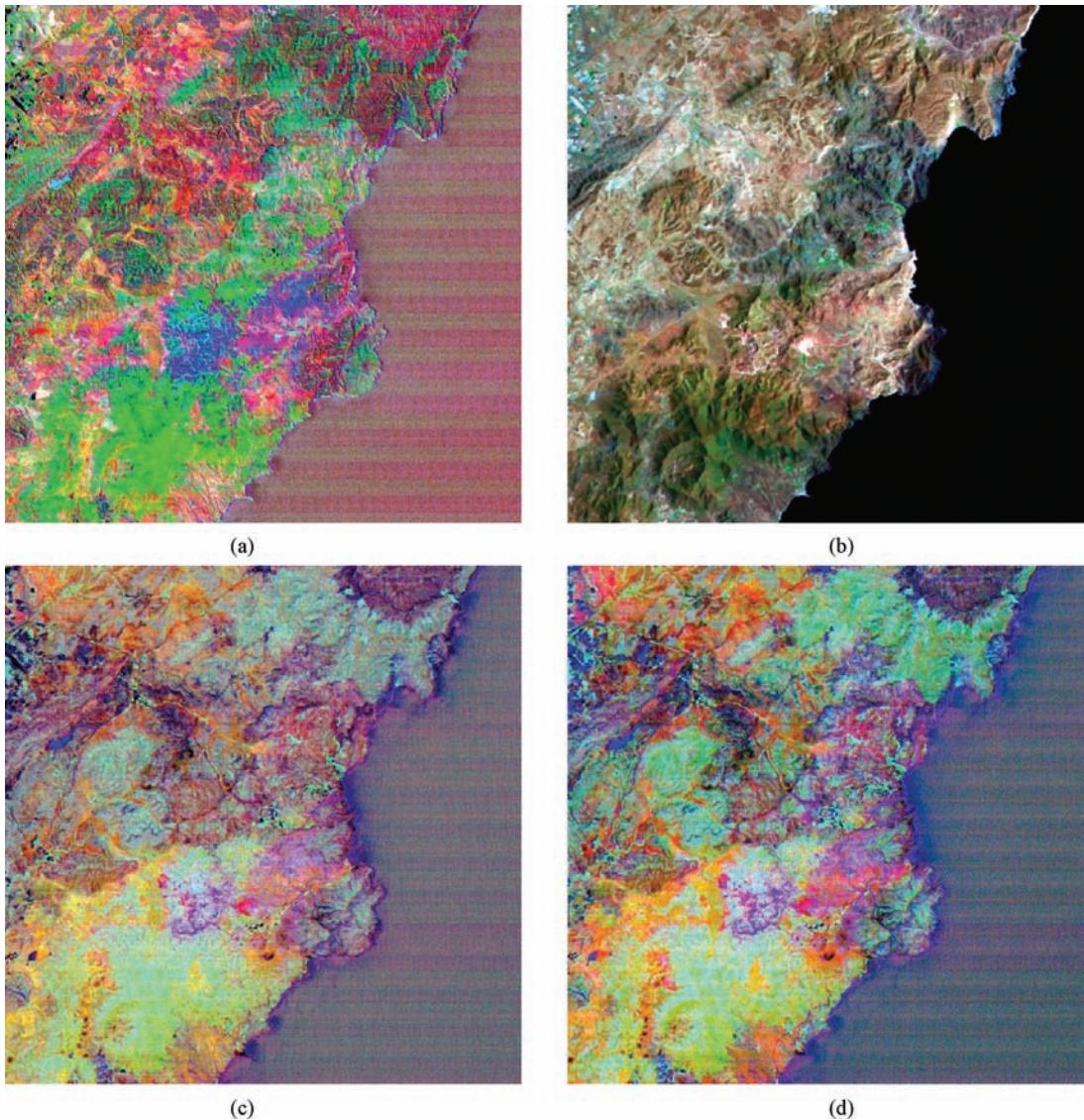
topography. Higher rank PCs contain significantly less information but it is more relevant to the spectral signatures of specific ground objects. Colour composites of PCs are often very effective for highlighting such ground objects and minerals which may not distinguishable in colour composites of the original bands. In PC colour composites noise may be exaggerated because the high rank PCs contain significantly less information than lower rank PCs and they have very low SNRs. When PC images are stretched and displayed in the same value range, the noise in higher rank PCs is improperly enhanced.

We would like to use three PCs with comparable information levels for colour composite generation. Chavez (1989) introduced a general approach, referred to as selective principal component analysis (SPCA), to produce PC colour composites in which the maximum information of either topographic or spectral features is condensed and in which the information content from each PC displayed is better balanced. There are two types of SPCA: *dimensionality and colour confusion reduction* and *spectral contrast mapping*.

#### 7.3.1 Dimensionality and colour confusion reduction

The spectral bands of a multi-spectral image are arranged into three groups and each group is composed of highly correlated bands. PCA is performed on each group and then the three PC1s derived from these three groups are used to generate an RGB colour composite. As the bands in each group are highly correlated, PC1 concentrates the maximum information of each group. For six reflective spectral bands of a TM or ETM+ image, the technique may condense more than 98% variance (information) in the derived PC1 colour composite. The recommended groups for six reflective spectral bands of TM or ETM+ images are given in Table 7.4.

The approach sounds clever but it is actually equivalent to generating a colour composite using broader spectral bands, given that PC1 is a positively weighted summation of the bands involved in the PCA. As illustrated in Figure 7.3b, it is essentially a colour composite of a broad visible band in blue, an NIR band in green and a broad SWIR band in red.



**Figure 7.3** PC colour composites derived from a TM sub-scene image: (a) colour composite of PC4 in red, PC3 in green and PC5 in blue; (b) SPCA colour composite using PC1s derived from band groups listed in Table 7.4; (c) SPCA (spectral contrast mapping) colour composite using PC2s derived from band groups listed in Table 7.5; (d) FPCS spectral contrast mapping using PC2s and PC3 derived from band groups listed in Table 7.6

### 7.3.2 Spectral contrast mapping

A more interesting and useful approach is spectral contrast mapping, where the primary objective is to map the contrast between different parts of the spectrum, and so to identify information unique to each band rather than information in common. For this purpose, PC2s derived from band pairs are used

instead of PC1s. By using only two bands as inputs, the information that is common to both bands is mapped to PC1 and the unique information is mapped to PC2. In general, low or medium correlation between the bands in each pair is preferred for this approach. The recommended grouping for the six reflective spectral bands of TM/ETM+ is listed in Table 7.5 as an example.

**Table 7.4** Dimensionality and colour confusion reduction for TM or ETM+

Groups	PCA	Colour
<i>TM1, 2, 3</i>	$PC1_{1,2,3}$	Blue
<i>TM4</i>		Green
<i>TM5, 7</i>	$PC1_{5,7}$	Red

Based on the above principle, groups of three or more bands may also be used for spectral contrast mapping. The technique can generate spectrally informative colour composites with significantly reduced topographic shadow effects, as illustrated in Figure 7.3c. In this colour composite, PC2 from bands 1 and 3 shows red spectral contrast relating to red soils and iron oxides, PC2 from bands 2 and 4 relates to vegetation, while PC2 from bands 5 and 7 is effectively highlighting the spectral contrast of clay alteration minerals. The striped pattern in the sea in this image is less obvious than that in the PC435 RGB colour composite in Figure 7.3a, implying better SNR.

### 7.3.3 FPCS spectral contrast mapping

The outcome of the spectral contrast mapping largely depends on the spectral band groupings. Knowing the spectral signatures of intended targets, we can use the FPCS method to decide the grouping of bands and thus the selection of PCs for the final RGB display. We demonstrate the principle of this approach with the same example as above.

From the eigenvector matrix in Table 7.3, we see that none of PCs picks up the ‘red-edge’ feature diagnostic of vegetation or the absorption feature of

**Table 7.5** TM or ETM+ spectral bands grouping for spectral contrast mapping

Groups	PCA	Colour
<i>TM1, 3</i>	$PC2_{1,3}$	Red
<i>TM2, 4</i>	$PC2_{2,4}$	Green
<i>TM5, 7</i>	$PC2_{5,7}$	Blue

**Table 7.6** FPCS spectral contrast mapping for TM/ETM+

Groups	Intended targets	FPCS	Colour
<i>TM1, 2, 3</i>	Red soils and iron oxide	$PC2_{1,2,3}$	Red
<i>TM2, 3, 4</i>	Vegetation	$PC2_{2,3,4}$	Green
<i>TM3, 5, 7</i>	Clay minerals	$PC3_{3,5,7}$	Blue

clay minerals characterized by the difference between *TM5* and *TM7*. We therefore consider the grouping listed in Table 7.6. From the eigenvector matrices in Table 7.7, we make the following observations:

1. PC2 derived from TM bands 1, 2 and 3 is essentially the difference between red (*TM3*) and blue (*TM1*) and it therefore enhances red features like iron oxides. This PC2 is chosen to display in red.
2. PC2 derived from TM bands 2, 3 and 4 is dominated by the positive contribution of NIR (*TM4*) and balanced by a negative contribution from red (*TM3*) and green (*TM2*). It therefore highlights healthy vegetation. This PC2 is chosen to display in green.
3. PC3 derived from TM bands 3, 5 and 7 is a summation of *TM3* and *TM5*, subtracting *TM7*, and therefore enhances clay alteration minerals and iron oxides. This PC3 is chosen to display in blue.

**Table 7.7** Eigenvector matrices of the three band groups for FPCS spectral contrast mapping

Eigenvector	PC1	PC2	PC3
<i>TM1</i>	0.507	<b>-0.834</b>	-0.218
<i>TM2</i>	0.458	<b>0.046</b>	0.888
<i>TM3</i>	0.731	<b>0.549</b>	-0.405
<i>TM2</i>	0.366	<b>-0.510</b>	-0.778
<i>TM3</i>	0.593	<b>-0.517</b>	0.617
<i>TM4</i>	0.717	<b>0.687</b>	-0.113
<i>TM3</i>	0.331	0.920	<b>0.210</b>
<i>TM5</i>	0.843	-0.388	<b>0.372</b>
<i>TM7</i>	0.424	0.054	<b>-0.904</b>

The resultant FPCS spectral contrast mapping colour composite in Figure 7.3d resembles the simple SPCA spectral contrast mapping colour composite in Figure 7.3c but the signatures of red soils/regoliths, vegetation and clay minerals are more distinctively displayed in red, green and blue.

After all the effort of SPCA, Figure 7.3 indicates that both the spectral contrast mapping and the FPCS spectral contrast mapping images are less colourful than the simple colour composite of PCs. One of the reasons for this is that the selected PCs from the three different band groups are not independent. They may be well correlated even though the PCs within each group are independent of each other. For instance, if we group TM bands 1 and 5 as one group and bands 2 and 7 as another, then the PC2s derived from the two groups will be highly correlated because both bands 1 and 2 and bands 5 and 7 are highly correlated. The way that the image bands are grouped will control the effectiveness of spectral contrast mapping.

## 7.4 Decorrelation stretch

A very important application of PCA is decorrelation stretch (DS). We have already learnt about two saturation-stretch-based DS techniques in Chapter 5, but the initial concept of the DS, as proposed by Taylor (1973), was based on PCA and further developed by Soha and Schwartz (1978).

The interpretation of PC colour composites is not straightforward and ordinary colour composites of the original image bands are often needed for reference. The PCA-based DS (PCADS) generates a colour composite from three image bands with reduced inter-band correlation and thus presents image spectral information in more distinctive and saturated colours without the distortion of hues. The idea of PCADS is to stretch multi-dimensional image data along their PC axes (the axes of the data ellipsoid cluster) rather than the original axes representing image bands. In this way, the volume of a data cluster can be effectively increased and the inter-band correlation is reduced as illustrated in the 2D case in Figure 7.4. PCADS is achieved in three steps:

1. PCA to transform the data from the original image bands to PCs.

2. Contrast enhancement on each of the PCs (stretching the data cluster along PC axes).
3. Inverse PCA to convert the enhanced PCs back to the corresponding image bands.

According to the PCA defined in Equation (7.6), inverse PCA is defined as

$$\Sigma_x = \mathbf{G}^{-1} \Sigma_y (\mathbf{G}^T)^{-1}. \quad (7.12)$$

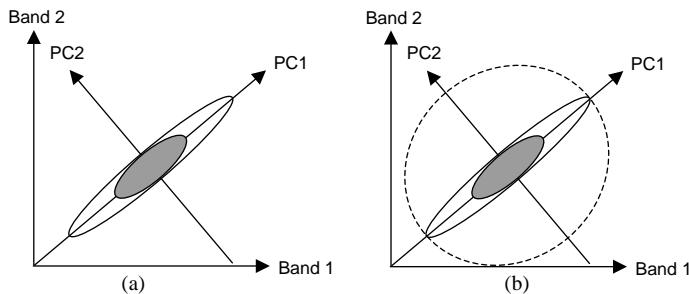
The PCADS technique effectively improves the colour saturation of a colour composite image without changing its hue characteristics. It is similar to IHSDS in its result but is based on quite different principles. PCADS is statistically scene dependent as the whole operation starts from the image covariance matrix, and it can be operated on all image bands simultaneously. IHSDS, in contrast, is not statistically scene dependent and only operates on three bands. Both techniques involve complicated forward and inverse coordinate transformations. In particular, PCADS requires quite complicated inverse operations on the eigenvector matrix and is therefore not computationally efficient. The direct decorrelation stretch (DDS) is the most efficient technique and it can be quantitatively controlled based on the saturation level of the image.

## 7.5 Physical-property-orientated coordinate transformation and tasseled cap transformation

Image PCA is a rotational operation of an  $m$ -dimensional orthogonal coordinate system for an  $m$ -band multi-spectral image  $\mathbf{X}$ . The rotation is scene dependent and determined by the eigenvector matrix  $\mathbf{G}$  of the covariance matrix  $\Sigma_x$ . Consider the rotation transform defined in Equation (7.3); in a general sense, we can arbitrarily rotate the  $m$ -dimensional orthogonal coordinate system in any direction as defined by a transformation  $\mathbf{R}$ :

$$\mathbf{y} = \mathbf{Rx}. \quad (7.13)$$

Here  $\mathbf{y}$  is a linear combination of  $\mathbf{x}$  specified by the coefficients (weights) in  $\mathbf{R}$ .



**Figure 7.4** Illustration of the decorrelation stretch in a 2D case: (a) stretching of the original bands is equivalent to stretching along the PC1 axis to elongate the data cluster; and (b) stretching along both the PC1 and PC2 axes expands the elliptical data cluster in two dimensions as denoted by the dashed line ellipse

For example, a 3D rotation from  $\mathbf{x}(x_1, x_2, x_3)$  to  $\mathbf{y}(y_1, y_2, y_3)$  is defined as

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} \cos \alpha_1 & \cos \beta_1 & \cos \gamma_1 \\ \cos \alpha_2 & \cos \beta_2 & \cos \gamma_2 \\ \cos \alpha_3 & \cos \beta_3 & \cos \gamma_3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \quad (7.14)$$

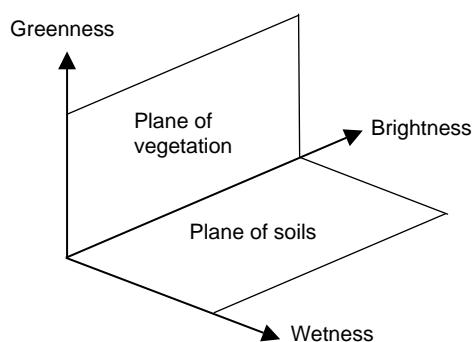
where subscript 1 denotes the rotation angles between the  $y_1$  axis and the  $x_1, x_2, x_3$  axes in the positive direction; subscript 2 the rotation angles between the  $y_2$  axis and the  $x_1, x_2, x_3$  axes in the positive direction, and so on for subscript 3.

In addition to the rotation, we can also consider a coordinate shift, as defined by a shift vector  $\mathbf{C}$ , and thus Equation (7.13) is modified as

$$\mathbf{y} = \mathbf{Rx} + \mathbf{C}. \quad (7.15)$$

Based on image data analysis of spectral signatures of particular targets, we can learn the data distribution of specific physical properties of ground objects, such as vegetation greenness, soil brightness and land surface wetness. We can then rotate the original image coordinate system to orientate to the directions of the maximum variation of these physical properties if these properties are orthogonal. This operation is quite similar to PCA but the rotation is decided by the transformation  $\mathbf{R}$  as derived from sample image data representative of the intended physical properties, rather than the eigenvector matrix  $\mathbf{G}$  which is derived from the image covariance matrix. As such, the rotational transformation  $\mathbf{R}$  is invariant to the images taken by the same multi-spectral sensor system on the one hand, but is constrained and biased by its empirical nature on the other hand.

One of the most successful examples of the physical-property-orientated coordinated transformation is the *tasseled cap transformation* initially derived by Kauth and Thomas (1976) for the Landsat MSS and then further developed by Crist and Cicone (1984) for the Landsat TM. As shown in Figure 7.5, the goal of the tasseled cap transformation is to transform the six reflective spectral bands (1–5 and 7) of TM or ETM+ in VNIR and SWIR spectral ranges, into three orthogonal components orientated as three key properties of the land surface: *brightness*, *greenness* (vigour of green vegetation) and *wetness*. The axes of brightness and greenness define the plane of vegetation presenting the 2D scattering of vegetation of varying greenness and grown on soils of different brightness, while the axes of wetness and brightness define the plane of soil presenting the 2D scattering of soil brightness in relation to soil moisture. Based mainly on a TM image of North Carolina taken on 24 September 1982 together with several other TM



**Figure 7.5** The tasseled cap transformation coordinate system. Adapted from Mather (2004)

scenes and simulated TM images, Crist and Cicone derived the TM tasseled cap transformation as

$$\begin{pmatrix} \text{Brightness} \\ \text{Greenness} \\ \text{Wetness} \end{pmatrix} = \begin{pmatrix} 0.3037 & 0.2793 & 0.4343 & 0.5585 & 0.5082 & 0.1863 \\ -0.2828 & -0.2435 & -0.5436 & 0.7243 & 0.0840 & -0.1800 \\ 0.1509 & 0.1793 & 0.3299 & 0.3406 & -0.7112 & -0.4572 \end{pmatrix} \begin{pmatrix} TM1 \\ TM2 \\ TM3 \\ TM4 \\ TM5 \\ TM7 \end{pmatrix}. \quad (7.16)$$

The transformation in (7.16) can also be expressed as in Table 7.8. Obviously, as a positive weighted summation of all the bands, the brightness is equivalent to a PC1 image. From an image with vegetation and soils at various wetness levels as the dominant land cover, we may locate higher rank PCs equivalent to greenness and wetness using the FPCS approach. However, for a multi-spectral image of a barren region where rock and regolith are dominant, the data variance representing vegetation will be very limited and not recognized as an axis of the data ellipsoid cluster. As a result, none of the PCs can represent greenness but using the tasseled cap transformation, the physical properties of greenness will be shown with a low and narrow value range. The tasseled cap transformation ‘pinpoints’ the predefined physical property and is

therefore scene independent and not affected by variation in land cover.

The tasseled cap transformation is widely used for its high relevance to the crop growth cycle, soil properties and surface moisture (both soil and vegetation). Since it is derived from sample TM image data of particular areas, and although quite representative in general terms, it is not a universal model nor is it correct for all the regions of the Earth; caution and critical assessment must be applied when using it. Considering variations in imaging conditions and environments, several variants of the tasseled cap transformation have been proposed. One example is the tasseled cap transformation for Landsat-7 ETM+ at-satellite reflectance, as shown in Table 7.9 (Huang *et al.*, 2002). It was derived by considering that effective atmospheric correction is often not feasible for regional applications.

**Table 7.8** Crist and Cicone’s TM tasseled cap transformation coefficients

TM band	1	2	3	4	5	7
Brightness	0.3037	0.2793	0.4343	0.5585	0.5082	0.1863
Greenness	-0.2828	-0.2435	-0.5436	0.7243	0.0840	-0.1800
Wetness	0.1509	0.1793	0.3299	0.3406	-0.7112	-0.4572

**Table 7.9** Landsat-7 ETM+ at-satellite reflectance tasseled cap transformation coefficients

ETM+ band	1	2	3	4	5	7
Brightness	0.3561	0.3972	0.3904	0.6966	0.2286	0.1596
Greenness	-0.3344	-0.3544	-0.4556	0.6966	-0.0242	-0.2630
Wetness	0.2626	0.2141	0.09926	0.0656	-0.7629	-0.5388

## 7.6 Statistic methods for band selection

With the increase in spectral resolution of remotely sensed data, more bands are available for colour composition, but only three bands can be used. For any particular application, it is not practical (or necessary) to exhaust all the possible three-band combinations for colour composition; band selection therefore becomes a quite important issue.

Band selection techniques can be divided in two major types: *statistical band selection* and *target-oriented band selection* techniques. Statistical techniques are normally used for selecting a few bands that may produce relatively high-quality colour composites of optimized visualization for general purposes. Target-oriented techniques are applied to highlight image features of particular interest. For instance, for vegetation studies, image bands in the VNIR spectrum are essential, while for mineral mapping, SWIR bands are the most effective.

For 7-band TM images, 11-band ATM images and 14-band ASTER images, statistical techniques are very effective tools to help users produce a few of the most informative colour composites from a large number of potential three-band combinations. However, for image data with much higher spectral resolution, such as AVIRIS (224 bands), the statistical approach becomes unfeasible. The band selection, in such cases, should be based on the spectral signatures of the particular targets. A combined approach could also be considered; that is, locating the relevant spectral range(s) from target spectral signatures first, then applying statistical techniques to select the most informative three-band groups for colour composition.

This section introduces a statistical band selection technique compiled by Crippen (1989) which is a more sound technique than the two widely used techniques briefly reviewed here.

### 7.6.1 Review of Chavez *et al.*'s and Sheffield's methods

Generally, bands with low inter-band correlation contain more information than highly correlated bands and therefore tend to produce colourful colour composites. Based on this principle, two band

selection techniques were developed by Chavez, Berlin and Sowers (1982) and Sheffield (1985).

Chavez *et al.*'s technique, the optimal index factor (OIF), was initially designed for selecting Landsat MSS ratios for ratio colour composites and was applied later to more advanced multi-spectral imagery (e.g. Landsat TM and ATM) for band selection. The OIF is defined as

$$\text{OIF} = \frac{SD_i + SD_j + SD_k}{|r_{ij}| + |r_{ik}| + |r_{jk}|} \quad (7.17)$$

where  $SD_i$  is the standard deviation for band  $i$  and  $r_{ij}$  is the correlation coefficient between band  $i$  and band  $j$ .

The largest OIF is considered to indicate the best band triplet for colour composition.

Sheffield's method is more general, and for selecting  $n$  bands from an  $m (>n)$  band image. The selection is based on the volume of the  $n$ -dimensional ellipsoids, as defined by  $n \times n$  principal sub-matrices of the covariance matrix, where a larger volume indicates higher information content. For colour composition, the data distributions of band triplets ( $n=3$ ) are represented by  $3 \times 3$  principal sub-matrices in 3D ellipsoids. The band triplet having an ellipsoid of the maximum volume is considered to contain the maximum information. Since the volume of the ellipsoid representing the data of a band triplet is decided by the value of the determinant of the correspondent 3D principal sub-matrix, selection is performed by computing and ranking the determinants of each  $3 \times 3$  principal sub-matrix of the  $m$ -band covariance matrix.

In both methods, the variances of image bands are considered as indicators of information content and are used as the basis for band selection. In fact, the variance of an image can be easily changed by contrast enhancement. A linear stretch does not change the information of an image and it is a common case for sensor calibration using gain factors, but it can affect band selection using these two techniques because a linear stretch increases the standard deviation in Chavez *et al.*'s method and the volume of the ellipsoid defined by the  $3 \times 3$  principal sub-matrices of the covariance matrix in Sheffield's method. A good band selection technique should be unaffected by linear stretch at the very least.

### 7.6.2 Index of three-dimensionality

If we leave aside the issue of image information content, we should realize that the main reason for a poor colour composite, that is a less colourful and informative one, is not the lack of information in each band but the information redundancy caused by high correlation between the three bands, as illustrated in Table 7.1. An effective statistical band selection technique should therefore aim to choose band triplets with minimum inter-band correlation and therefore minimum information redundancy.

We can prove that the correlation coefficients among different bands are independent of linear operations. Let  $r_{ij}$  represent the correlation coefficient between two image bands,  $X_i$  and  $X_j$ ; then

$$r_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}} \quad (7.18)$$

where  $\sigma_{ii}$  and  $\sigma_{jj}$  are the variances of band  $i$  and  $j$  and  $\sigma_{ij}$  is the covariance between the two bands.

If band  $X_i$  is enhanced by a linear stretch,  $Y_i = a_i X_i + b_i$ , and for band  $X_j$ ,  $Y_j = a_j X_j + b_j$ . Then the variances and covariance of  $Y_i$  and  $Y_j$  are

$$a_i^2 \sigma_{ii} \quad a_j^2 \sigma_{jj} \quad \text{and} \quad a_i a_j \sigma_{ij}$$

respectively. Thus the correlation coefficient between  $Y_i$  and  $Y_j$  is

$$R_{ij} = \frac{a_i a_j \sigma_{ij}}{\sqrt{a_i^2 \sigma_{ii} a_j^2 \sigma_{jj}}} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}} = r_{ij}. \quad (7.19)$$

Therefore, a band selection technique based on band correlation coefficients is not affected by linear contrast enhancement.

Based on this principle, Crippen (1989) proposed to use the square root of the determinant of the three-band correlation matrix as a measurement of the three-dimensionality of the three-band data distribution. A high three-dimensionality indicates a spherical data distribution and so a low inter-band correlation. The determinant of the correlation matrix of band  $i$ ,  $j$  and  $k$  is

$$\begin{vmatrix} r_{ii} & r_{ij} & r_{ik} \\ r_{ji} & r_{jj} & r_{jk} \\ r_{ki} & r_{kj} & r_{kk} \end{vmatrix} = 1 + 2r_{ij}r_{ik}r_{jk} - r_{ij}^2 - r_{ik}^2 - r_{jk}^2.$$

The *three-dimensionality index* for band selection is thus defined as

$$3D \text{ index} = \sqrt{1 + 2r_{ij}r_{ik}r_{jk} - r_{ij}^2 - r_{ik}^2 - r_{jk}^2}. \quad (7.20)$$

The value range of the 3D index is [0,1], where 1 indicates perfect three-dimensionality and 0 no three-dimensionality at all. The higher the index value, the better the statistical choice of a band triplet for colour composition.

### 7.7 Remarks

It is interesting to note that, although PCA is based on quite complex covariance matrix operations, in the end a PC image is simply a linear combination of the original image bands. In analysing the eigenvector of a PC image, with the FPCS technique, we are essentially selecting the PC through image differencing. High-rank PCs are nothing more than compound difference images but these are so composed as to be independent of one another. PCA ensures orthogonal (independent) PCs on the basis of the data distribution, while differencing allows the targeting of specific spectral signatures of interest, although the resulting difference images are not themselves orthogonal. FPCS combines the merits of both of these but may not always reveal the diagnostic spectral features.

Also noteworthy is that the average image, IHS intensity image and PC1 image share a great deal in common. The three images all represent the sum of the spectral bands and all increase the image SNR. A band average is an equal weight summation of any number of image bands, the IHS intensity image is an average of three bands used for the RGB–IHS transformation, while PC1 is a weighted summation of all image bands involving PCA.

The concept of DS is rooted in PCA but the PCADS is less efficient and less widely used than the saturation DS techniques because it involves more complicated matrix operations in the inverse PC transformation. Although the two types of DS technique are based on different principles, their effects on the bands of an RGB colour composite triplet are entirely equivalent: they both increase the three-dimensionality of the data cluster.

## Questions

- 7.1 Using a diagram, explain the principle of PCA.
- 7.2 Discuss the data characteristics of PC images and their applications.
- 7.3 Compare and contrast the images of band average, IHS intensity and PC1, and discuss their relative merits.
- 7.4 What are the major advantages and disadvantages of PC colour composition?
- 7.5 Describe the feature-oriented PC selection (FPCS) method and discuss its application to PC colour composition.
- 7.6 Discuss the two SPCA techniques and their applications.
- 7.7 Describe and comment on the combined approach of FPCS and SPCA for spectral enhancement.
- 7.8 What is a decorrelation stretch (DS)? Describe the major steps of PCADS.
- 7.9 Compare the three DS techniques, PCADS, IHSDS and DDS, in principle, results and processing efficiency.
- 7.10 What do PCA and the physical-property-orientated coordinate transformation have in common? How are they different in methodology and applications?
- 7.11 In what sense are the PCA and the tasseled cap transformation similar? What is the major difference between the two techniques? Comment on the merits and drawbacks of the two methods for Earth observation.
- 7.12 Describe the two main approaches for band selection. Why is band selection necessary for visualization of multi-spectral images?
- 7.13 Describe the principles behind the index of three-dimensionality. What is the major consideration in the design of this technique?
- 7.14 From the correlation matrix in Table 7.1, calculate the index of three-dimensionality using formula (7.20) for band triplets 321, 432, 531 and 541.

# 8

## Image Classification

Image classification belongs to a very active field in computing research, that of *pattern recognition*. Image pixels can be classified either by their multi-variable statistical properties, such as the case of multi-spectral classification (clustering), or by segmentation based on both statistics and spatial relationships with neighbouring pixels. In this chapter, we will look at multi-variable statistical classification techniques for image data.

### 8.1 Approaches of statistical classification

Generally, statistical classification can be catalogued into two major branches: *unsupervised* and *supervised classifications*.

#### 8.1.1 Unsupervised classification

This is entirely based on the statistics of the image data distribution, and is often called *clustering*. The process is automatically optimized according to cluster statistics without the use of any knowledge-based control (i.e. ground truth). The method is therefore objective and entirely data driven. It is particularly suited to images of targets or areas where there is no ground truth knowledge or where such information is not available, such as in the

case of planetary images. Even for a well-mapped area, unsupervised classification may reveal some spectral features which were not apparent beforehand. The result of an unsupervised classification is an image of statistical clusters, where the thematic contents of the clusters are not known. Ultimately, such a classification image still needs interpretation based on some knowledge of ground truth.

#### 8.1.2 Supervised classification

This is based on the statistics of *training areas* representing different ground objects selected subjectively by users on the basis of their own knowledge or experience. The classification is controlled by users' knowledge but, on the other hand, is constrained and may even be biased by their subjective view. The classification can therefore be misguided by inappropriate or inaccurate training area information and/or incomplete user knowledge.

Realizing the limitations of both major classification methods, a *hybrid classification* approach has been introduced. In the hybrid classification of a multi-spectral image, firstly an unsupervised classification is performed, then the result is interpreted using ground truth knowledge and, finally, the original image is reclassified using a supervised classification with the aid of the statistics of the

unsupervised classification as training knowledge. This method utilizes unsupervised classification in combination with ground truth knowledge as a comprehensive training procedure and therefore provides more objective and reliable results.

### 8.1.3 Classification processing and implementation

A classification may be completed in one step, as a *single pass classification*, or in an iterative optimization procedure referred to as an *iterative classification*. The single pass method is the normal case for supervised classification while the iterative classification represents the typical approach to unsupervised classification (clustering). The iterative method can also be incorporated into a supervised classification algorithm.

Most image processing software packages perform image classification in the image domain by *image scanning classification*. This approach can classify very large image datasets with many spectral bands and very high quantization levels, with very low demands on computing resources (e.g. RAM) but it cannot accommodate sophisticated classifiers (decision rules). Image classification can also be performed in feature space by *feature space partition*. In this case, sophisticated classifiers incorporating data distribution statistics can be applied but the approach demands a great deal of computer memory to cope with the high data dimensionality and quantization. This problem is being overcome by increasingly powerful computing hardware and dynamic memory management in programming.

### 8.1.4 Summary of classification approaches

These are as follows:

- Unsupervised classification
- Supervised classification
- Hybrid classification
- Single pass classification
- Iterative classification
- Image scanning classification
- Feature space partition.

## 8.2 Unsupervised classification (iterative clustering)

### 8.2.1 Iterative clustering algorithms

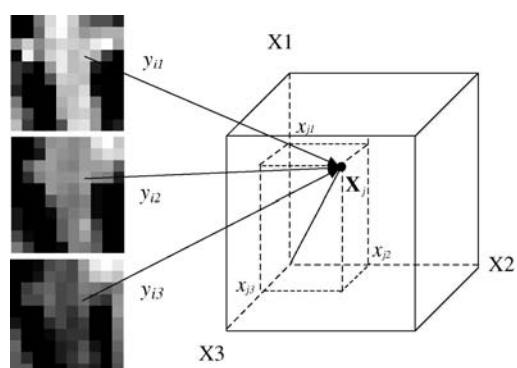
For convenience of description, let  $\mathbf{X}$  be a  $n$ -dimensional feature space of  $n$  variables ( $x_1, x_2, \dots, x_n$ ),  $Y_i$  be an object of an object set  $Y$  (an image) defined by measurements of the  $n$  variables (e.g. DNs of  $n$  spectral bands),  $Y_i = (y_{i1}, y_{i2}, \dots, y_{in})$ ,  $i = 1, 2, \dots, N$ .  $N$  is the total number of objects in  $Y$  or the total number of pixels in an image. As shown in Figure 8.1, in the feature space  $\mathbf{X}$ , the object  $Y_i$  is represented by an observation vector, that is a data point  $\mathbf{X}_j \in \mathbf{X}$  at the coordinates  $(x_{j1}, x_{j2}, \dots, x_{jn})$ ,  $j = 1, 2, \dots, M$ .  $M$  is the total number of data points representing  $N$  objects. If  $\mathbf{X}$  is a Euclidean space, then  $x_{jh} \sim y_{ih}$ ,  $h = 1, 2, \dots, n$ . Obviously, a data point  $\mathbf{X}_j$  in the feature space  $\mathbf{X}$  can be shared by more than one image pixel  $Y_i$  and therefore  $M \leq N$ .

The goal of the clustering process is to identify the objects of set  $Y$  in  $m$  classes. This is equivalent to the partition of the relevant data points in feature space  $\mathbf{X}$  into  $m$  spatial clusters,  $\omega_1, \omega_2, \dots, \omega_m$ . Generally, there are two principal iterative clustering algorithms labelled  $\alpha$  and  $\beta$  (Diday and Simon, 1976), as follows:

#### Algorithm $\alpha$

##### 1. Initialization

Let  $m$  elements  $Y_q \in Y$ , chosen at random or by a selection scheme, be the ‘representation’



**Figure 8.1** A 3D illustration of the relationship between a feature space point  $\mathbf{X}_j$  and a multi-spectral image pixel  $Y_i = (y_{i1}, y_{i2}, y_{i3})$

- of  $m$  clusters denoted as  $\omega_1, \omega_2, \dots, \omega_k, \dots, \omega_m$ .
2. *Clustering*  
For all  $i$ , assign any element  $Y_i (Y_i \in Y)$  to a cluster  $\omega_k$ , if the dissimilarity measurement  $\delta(Y_i, \omega_k)$  is minimal.
  3. *Update statistical representation*  
For all  $k$ , new statistics of cluster  $\omega_k$  are computed as the renewed representation of the cluster  $\omega_k$ .
  4. *Stability*  
If no  $\omega_k$  has changed above the given criteria then stop, else go to 2.

### Algorithm $\beta$

1. As in step 1 of algorithm  $\alpha$ .
2. One element  $Y_i (Y_i \in Y)$  is assigned to cluster  $\omega_k$ , if  $\delta(Y_i, \omega_k)$  is minimal.
3. A new representation of  $\omega_k$  is computed from all the elements of cluster  $\omega_k$ , including the last element.
4. If all elements  $Y_i (Y_i \in Y)$  have been assigned to a cluster then stop, else go to step 2.

Algorithm  $\alpha$  may not necessarily converge if the criterion for terminating the iteration is too tight. Algorithm  $\beta$  ends when the last pixel is reached. Algorithm  $\alpha$  is more commonly used for image classification because of its self-optimization mechanism and processing efficiency. Cluster splitting and merging functions can be added to algorithm  $\alpha$  after step 4, which allows the algorithm to operate more closely with the true data distribution and to reach more optimized convergence. During the progress of the clustering iteration, the initial cluster centres move towards the true data cluster centres via the updating of their statistical representations at the end of each iteration. The only user control on clustering is the initial parameter setting, such as number and position of the starting centres of clusters, iteration times or termination criteria, maximum and minimum number and size of clusters, and so on. The initial setting will affect the final result. In this sense, the clustering iteration mechanism can only ensure local optimization, the optimal partition of clusters for the given initial parameter setting, but not the global optimization, because the initial parameter setting cannot be optimal for the best possible clustering result.

For most image processing packages, image clustering using either of the two algorithms is executed on an object set  $Y$ , that is the image. The processing is on a pixel-by-pixel basis by scanning the image but, with advances in computing power, the very demanding feature space partition clustering in the feature space  $\mathbf{X}$  becomes feasible.

One of the most popular clustering algorithms for image classification, the ISODATA algorithm (Ball, 1965; Ball and Hall, 1967), is a particular case of algorithm  $\alpha$  in which the dissimilarity measure  $\delta(Y_i, \omega_k)$  in step 2 is the square Euclidean distance. The assumption underlying this simple and efficient technique is that all the clusters have equal variance and population. This assumption is generally untrue in image classification, and as a result classification accuracy may be low. To improve ISODATA, more sophisticated measures of dissimilarity, such as maximum likelihood estimation and population weighted measurements, have been introduced. For all these different decision rules, within the ISODATA frame, the processing is performed by image scanning.

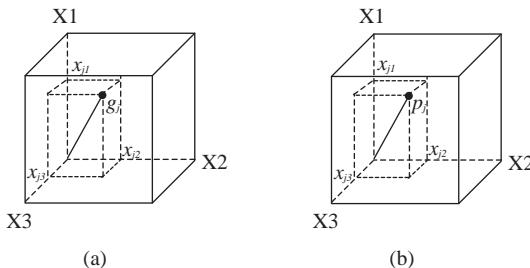
### 8.2.2 Feature space iterative clustering

As mentioned earlier, image classification can be performed by image scanning as well as by feature space partition. Most multi-variable statistical classification algorithms can be realized by either approach but, for more advanced decision rules, such as optimal multiple point reassignment (OMPR), which will be introduced later, feature space partition is the only feasible method because all pixels sharing the same DN values in each image band must be considered simultaneously. Here we introduce a three-dimensional feature space iterative clustering method, the 3D-FSIC method, an algorithm which can be easily extended to further dimensions.

#### Three-dimensional feature space iterative clustering (3D-FSIC)

Step 1. Create a 3D Scattergram of the Input Image  
Read the input image,  $Y$ , pixel by pixel and record the pixel frequencies in a 3D scattergram, that is a 3D array (Figure 8.2a)

$$G(d1 \times d2 \times d3)$$



**Figure 8.2** (a) A 3D array  $G$  of a scattergram; and (b) a 3D array  $P$  of a feature space partition

where  $d_1, d_2$  and  $d_3$  are the array sizes in the three dimensions or the maximum DN values of the three bands of the image  $Y$ .

The value of any element  $g_j$  in  $G$  indicates how many pixels share the point  $X_j$  at the coordinates  $(x_{j1}, x_{j2}, x_{j3})$  in the 3D feature space  $\mathbf{X}$ , or the number of pixels with the same DN values as pixel  $Y_i$  in the image  $Y$ , where  $y_{ih} \sim x_{jh}$ ,  $h = 1, 2, 3$ .

**Step 2. Initialization** Select  $m$  points in the 3D feature space  $\mathbf{X}$  as the ‘seeds’ of  $m$  clusters and call them  $\omega_k, k = 1, 2, \dots, m$ . The choice could be made at random or via an automatic seed selection technique.

**Step 3. Feature space clustering** For all  $j$ , assign any point  $\mathbf{X}_j$  ( $\mathbf{X}_j \in \mathbf{X}, j = 1, 2, \dots, N$ ) to cluster  $\omega_k$  if the dissimilarity  $\delta(\mathbf{X}_j, \omega_k)$  is minimal. Thus all the pixels sharing the point  $\mathbf{X}_j$  are assigned to cluster  $\omega_k$  simultaneously. The size of cluster  $\omega_k$ ,  $N_k$ , increases by the value  $g_j$  while, if it is a reassignment, the size of the cluster to which  $\mathbf{X}_j$  was formerly assigned decreases by the same value. The cluster sequential number  $k$  of point  $\mathbf{X}_j$  is recorded by a 3D feature space partition array  $P$  ( $d_1 \times d_2 \times d_3$ ) in the element  $p_j$  at coordinates  $(x_{j1}, x_{j2}, x_{j3})$  (Figure 8.2b).

**Step 4. Update the statistical representation of each cluster** For all  $k$  ( $k = 1, 2, \dots, m$ ), statistical parameters, such as mean vector  $\mu_k$ , covariance matrix  $\Sigma_k$  and so on are calculated. These parameters comprise the new representation of the cluster  $\omega_k$ .

**Step 5. Stability** For all  $k$  ( $k = 1, 2, \dots, m$ ), if the maximum spatial migration of the mean vector

$\mu_k$  (the kernel of the cluster) is less than a user-controlled criterion, go to step 7, else go to step 6.

**Step 6. Cluster splitting and merging** Split the overlarge and elongate clusters and merge clusters which are too small and/or too close to each other, according to user-controlled criteria; then update the statistical representations of the new clusters. Go to step 3.

**Step 7. Transfer the Clustering Result from Feature Space to an Image** Read the input image  $Y$ , pixel by pixel. For all  $i$ , assign a pixel  $Y_i$  ( $Y_i \in Y$ ) to cluster  $\omega_k$  if its relevant data point  $\mathbf{X}_i$  in feature space  $\mathbf{X}$  is assigned to this cluster, according to the record in the feature space partition array  $P$ , that is

$$Y_i \rightarrow \omega_k \text{ if } P_j = k$$

where  $P_j$  is at coordinates  $(x_{j1}, x_{j2}, x_{j3})$  in  $P$  and  $y_{ih} \sim x_{jh}$ ,  $h = 1, 2, 3$ .

Then assign the class number to the corresponding pixel in the output classification image  $Y_{\text{class}}$ .

### 8.2.3 Seed selection

The initial kernels (seeds) for unsupervised classification can be made randomly, evenly or by particular methods. Here we introduce an automatic seed selection technique (ASST) for 3D-FSIC.

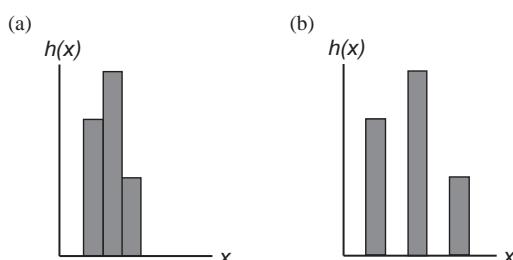
In the 3D scattergram of the three images for classification, data will exhibit peaks (the points of high frequency) at the locations of spectral clusters. It is thus sensible to use these points as the initial kernels of clusters to start iterative clustering. Such a peak point has two properties:

- Higher frequency than all its neighbouring points in the feature space.
- Relatively high frequency in the 3D scattergram.

These two properties are to be used to locate peak points. It is important to bear in mind that the multispectral image data and scattergram are discrete and that the DN value increment of an image may not necessarily be unity, especially after contrast enhancement. For instance, when an image of 7 bit

DN range [0, 127] is linearly stretched to 8 bit DN range [0, 255], the increment of DN values becomes 2 instead of 1. In this case, any non-zero frequency DN level in the original image will have two adjacent zero-frequency DN levels in the stretched image (Figure 8.3), appearing as a pseudo peak caused by data discontinuity. With these considerations in mind, ASST is composed of the following operations:

1. Locate and rank the first  $N$  points of highest frequency from the 3D scattergram to form a sequential set  $\mathbf{X}_c$ .  $N$  can be decided by setting a criterion frequency on the basis of experience and experimentation. For an 8 bit full-scene TM image, 1024 is suggested. This operation will prevent the selection of isolated low-frequency points (often representing noise) as seeds and thus satisfy the second property.
2. The first element in the set  $\mathbf{X}_c$  must be nominated as a seed because it cannot be surrounded by any elements of a higher frequency. Then, for the second element of  $\mathbf{X}_c$ , check if the first element is in its given neighbourhood range (the neighbourhood is used to avoid pseudo peaks in the image with DN increment greater than 1) and, if not, the second element is also selected as a seed. In general, for any element  $\mathbf{X}_j$  in  $\mathbf{X}_c$ , check the coordinates of those elements ranked with higher frequency;  $\mathbf{X}_j$  is selected as a seed if none of the higher frequency elements are within its neighbourhood. This operation makes the seed selection satisfy the first property.



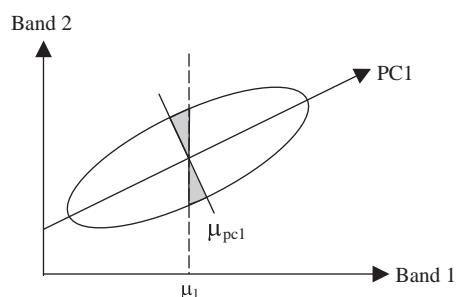
**Figure 8.3** Illustration of pseudo peaks in an image histogram (a) caused by linear stretch (b)

### 8.2.4 Cluster splitting along PC1

In unsupervised classification (cluster partition) very large clusters may be generated. Such large clusters may contain several classes of ground objects. A function for cluster splitting is therefore necessary to achieve optimal convergence. In ISODATA, an overlarge and elongated cluster  $\omega$  is split according to the variable with greatest standard deviation. The objects (image pixels) in cluster  $\omega$  are reassigned to either of the two new clusters,  $\omega_1$  and  $\omega_2$ , depending on whether their splitting variable values are above or below the mean of the splitting variable. As shown by the 2D case in Figure 8.4, splitting in this way may cause incorrect assignments of those objects in the shaded area of the data ellipse. They are assigned to a new cluster which is farther away from them rather than closer. This error can be avoided if the cluster  $\omega$  is split along its first principal component (PC1). Since PC1 can be found without performing a principal component transformation, not too many calculations are involved. The technique of cluster splitting based on PC1 (Liu and Haigh, 1994) includes two steps: finding PC1 followed by cluster splitting based on PC1.

#### 8.2.4.1 Find the first principal component PC1

The covariance matrix  $\Sigma$  of the cluster  $\omega$  is a non-negative definite matrix. Thus the first eigenvalue and eigenvector of  $\Sigma$ ,  $\lambda_1$  and  $\mathbf{a} = (a_1, a_2, \dots, a_n)^T$ ,



**Figure 8.4** A 2D example of cluster splitting based on band 1 (the variable with maximum standard deviation) and PC1. The shaded areas indicate the misclassification resulting from the cluster splitting based on band 1

can be found by the iteration

$$\begin{aligned}\Sigma \mathbf{a}^{(s)} &= \lambda_1^{(s+1)} \mathbf{a}^{(s+1)} \\ \mathbf{a}^{(0)} &= \mathbf{I}\end{aligned}\quad (8.1)$$

where  $s$  denotes the number of iterations and  $\mathbf{I}$  is an identity vector.

As an eigenvector,  $\mathbf{a}$  is orthogonal; thus for each iteration  $s$ , we have

$$(\mathbf{a}^{(s)})^T \mathbf{a}^{(s)} = 1. \quad (8.2)$$

Then

$$\begin{aligned}(\Sigma \mathbf{a}^{(s)})^T \Sigma \mathbf{a}^{(s)} &= \lambda_1^{(s+1)} (\mathbf{a}^{(s+1)})^T \lambda_1^{(s+1)} \mathbf{a}^{(s+1)} \\ &= (\lambda_1^{(s+1)})^2.\end{aligned}$$

Thus,

$$\begin{aligned}\lambda_1^{(s+1)} &= [(\Sigma \mathbf{a}^{(s)})^T \Sigma \mathbf{a}^{(s)}]^{1/2} \\ a^{(s+1)} &= \frac{\Sigma \mathbf{a}^{(s)}}{\lambda_1^{(s+1)}}.\end{aligned}\quad (8.3)$$

After 5–6 iterations convergence with an accuracy higher than  $10^{-5}$  can be achieved and the first eigenvalue  $\lambda_1$  and eigenvector  $\mathbf{a}$  are found. Consequently, the first principal component of cluster  $\omega$  in the  $n$ -dimensional feature space  $\mathbf{X}$  is derived as

$$\text{PC1} = (\mathbf{a})^T \mathbf{X} = \sum_{h=1}^n a_h \mathbf{x}_h. \quad (8.4)$$

#### 8.2.4.2 Cluster splitting

According to (8.4), the PC1 coordinate of the mean vector  $\boldsymbol{\mu} = (\mu_1, \mu_2, \dots, \mu_n)^T$  of cluster  $\omega$  is

$$\mu_{\text{pc1}} = (\mathbf{a})^T \boldsymbol{\mu} = \sum_{h=1}^n a_h \mu_h. \quad (8.5)$$

For every data point  $\mathbf{X}_j \in \omega$ , we calculate its PC1 coordinate:

$$x_{j,\text{pc1}} = (\mathbf{a})^T \mathbf{X}_j = \sum_{h=1}^n a_h x_{jh}. \quad (8.6)$$

We assign  $\mathbf{X}_j$  to  $\omega_1$  if  $x_{j,\text{pc1}} > \mu_{\text{pc1}}$ , otherwise we assign  $\mathbf{X}_j$  to  $\omega_2$ .

Cluster splitting can also be performed on the objects (image pixels) instead of data points by replacing  $\mathbf{X}_j$  by  $Y_i$  ( $i = 1, 2, \dots, N$ ) in formula (8.6).

After cluster splitting, the statistics of the two new clusters are calculated as the representations for the next clustering iteration.

## 8.3 Supervised classification

### 8.3.1 Generic algorithm of supervised classification

A supervised classification comprises three major steps, as follows:

**Step 1. Training** Training areas representing different ground objects are manually and interactively defined on the image display. Statistics of the training areas are calculated to represent the relevant classes  $\omega_k$  ( $k = 1, 2, \dots, m$ ).

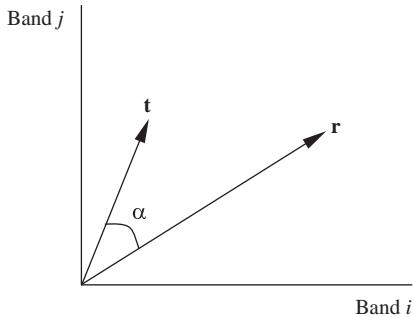
**Step 2. Classification** For all  $i$ , assign any element  $Y_i$  ( $Y_i \in Y$ ) to a class  $\omega_k$ , if the dissimilarity measurement  $\delta(Y_i, \omega_k)$  is minimal.

**Step 3. Class statistics** Calculate the statistics of all resultant classes.

Iteration and class splitting/merging functions can also be accommodated into a supervised classification algorithm to provide an automated optimization mechanism.

### 8.3.2 Spectral angle mapping classification

A pixel in an  $n$ -band multi-spectral image can be considered as a vector in the  $n$ -dimensional feature space  $\mathbf{X}$ . The magnitude (length) of the vector is decided by the pixel DNs of all the bands while the orientation of the vector is determined by the shape of the spectral profile of this pixel. If two pixels have similar spectral properties but are under different solar illumination because of topography, the vectors representing the two pixels will have different lengths but very similar orientation. Therefore the classification of image pixels based on the spectral angles between them will be independent of topography (illumination) as well as any unknown linear translation factors (e.g. gain and offset). The spectral angle mapping (SAM) technique, proposed by Kruse, Lefkoff and Dietz (1993), is a supervised classification based on the angles between image



**Figure 8.5** A 2D illustration of two spectral vectors and the spectral angle ( $\alpha$ ) between them

pixel spectra and training data spectra or library spectra. The algorithm determines the similarity between two spectra by calculating the spectral angle between them as shown in a 2D diagram (Figure 8.5). According to vector algebra, the angle between two vectors  $\mathbf{r}$  and  $\mathbf{t}$  is defined as

$$\alpha = \arccos\left(\frac{\mathbf{t} \cdot \mathbf{r}}{|\mathbf{t}| \cdot |\mathbf{r}|}\right) \quad (8.7)$$

or

$$\alpha = \arccos\left(\frac{\sum_{i=1}^m t_i r_i}{(\sum_{i=1}^m t_i^2)^{1/2} (\sum_{i=1}^m r_i^2)^{1/2}}\right). \quad (8.8)$$

where  $m$  is the number of spectral bands.

The value range of  $\alpha$  is  $0-\pi$ .

In general, for  $N$  reference spectral vectors  $\mathbf{r}_k$  ( $k = 1, 2, \dots, N$ ), either from an existing spectral library or from training areas, the spectral vector  $\mathbf{t}$  of an image pixel is identified as  $\mathbf{r}_k$  if the angle  $\alpha$  between them is minimal and is less than a given criterion.

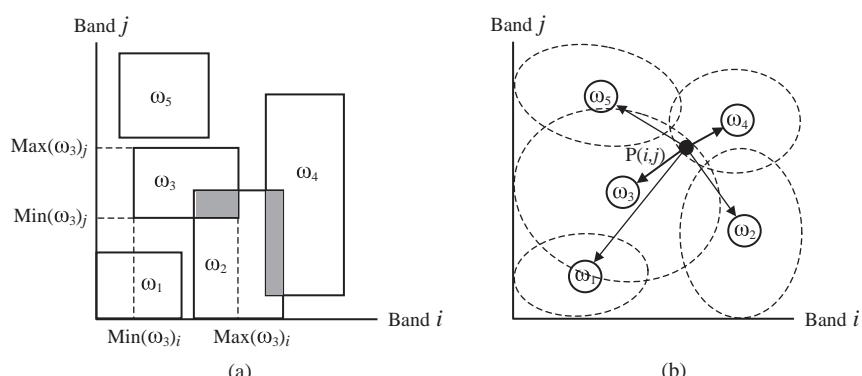
The SAM classification is widely used in hyperspectral image data classification for mineral identification and mapping. It can also be used in broadband multi-spectral image classification. Within the framework of SAM, different dissimilarity functions can be implemented to assess the spectral angle,  $\alpha$ .

## 8.4 Decision rules: dissimilarity functions

Dissimilarity functions, based on image statistics, formulate decision rules at the core of both supervised and unsupervised classification algorithms, and these theoretically decide the accuracy of a classification algorithm. Here we introduce several commonly used decision rules of increasing complexity.

### 8.4.1 Box classifier

This is also called a parallel classifier. It is used for single pass supervised classification. In principle, it is simply multi-dimensional thresholding (Figure 8.6a).



**Figure 8.6** Illustrations of 2D feature space partition of the box classifier and distance-based classifications: (a) a box classifier is actually a simple multi-dimensional threshold – it cannot classify image pixels that fall in the value ranges of multiple classes as shown in the shaded areas; (b) the circles are the class centres and the ellipses represent the size of each class. The minimum Euclidean distance classification will assign the pixel  $P(i,j)$  to the class centre  $\omega_4$  whereas the maximum likelihood minimum distance classification will be more likely to assign the pixel  $P(i,j)$  to the class centre  $\omega_3$  because this class is larger

For all  $i$ , assign an element  $Y_i$  ( $Y_i \in Y$ ) to cluster  $\omega_k$  if

$$\min(\omega_k) \leq Y_i \leq \max(\omega_k). \quad (8.9)$$

The ‘boxes’ representing the scopes of different classes may partially overlap one another, as in the shaded areas shown in Figure 8.6a. The pixels that fall within the overlap areas are treated as unclassified. This is a very crude but fast classifier.

#### 8.4.2 Euclidean distance: simplified maximum likelihood

The Euclidean distance is a special case of maximum likelihood which assumes equal standard deviation and population for all clusters. It is defined as follows.

For all  $i$ , assign an element  $Y_i$  ( $Y_i \in Y$ ) to cluster  $\omega_k$  if

$$d(Y_i, \omega_k) = (Y_i - \mu_k)^T (Y_i - \mu_k) = \min\{d(Y_i, \omega_r)\} \quad (8.10)$$

for  $r = 1, 2, \dots, m$  and where  $\mu_k$  is the mean vector of cluster  $\omega_k$ .

The Euclidean distance lies at the core of the ISODATA minimum distance classification.

#### 8.4.3 Maximum likelihood

The maximum likelihood decision rule is based on Bayes’ theorem and assumes a normal distribution for all clusters. In this decision rule, the feature space distance between an image pixel  $Y_i$  and cluster  $\omega_k$  is weighted by the covariance matrix  $\Sigma_k$  of  $\omega_k$  with an offset relating to the ratio of  $N_k$ , the number of pixels in  $\omega_k$ , to  $N$ , the total number of pixels of the image  $Y$ .

For all  $i$ , assign an element  $Y_i$  ( $Y_i \in Y$ ) to cluster  $\omega_k$  if

$$\begin{aligned} \delta(Y_i, \omega_k) &= \ln|\Sigma_k| + (Y_i - \mu_k)^T \Sigma_k^{-1} (Y_i - \mu_k) \\ -\ln \frac{N_k}{N} &= \min\{\delta(Y_i, \omega_r)\} \end{aligned} \quad (8.11)$$

for  $r = 1, 2, \dots, m$ .

As shown in Figure 8.6b, the minimum Euclidean distance classification will assign the pixel  $P(i, j)$  to the class centre  $\omega_4$ , whereas the maximum likelihood minimum distance classification will be more likely to assign the pixel  $P(i, j)$  to the class centre  $\omega_3$  because this class is larger.

#### 8.4.4 \*Optimal multiple point reassignment

An advantage of 3D-FSIC is that the optimal multiple point reassignment (OMPR) rule can be implemented if we let  $\delta(\mathbf{X}_j, \omega_k)$  be an OMPR dissimilarity measurement at step 3 of 3D-FSIC. The OMPR (Kittler and Pairman, 1988) was developed based on the optimal point assignment rule (Macqueen, 1967). By using OMPR, the cluster sizes and the number of pixels sharing the same data point in feature space (point frequency) are taken into account when a reassignment of these pixels is made. Thus the accuracy of the clustering partition can be reasonably improved.

Suppose a data point  $\mathbf{X}_j$  currently allocated to cluster  $\omega_l$  is shared by  $H$  pixels; then the OMPR based on the square Euclidean distance (Euclidean OMPR) for all these pixels from cluster  $\omega_l$  to cluster  $\omega_k$ , shared by  $N_k$  pixels, will be achieved if  $\omega_k$  satisfies

$$\begin{aligned} \frac{N_k}{N_k + H} d(\mathbf{X}_j, \mu_k) &= \min_{r \neq l} \frac{N_r}{N_r + H} d(\mathbf{X}_j, \mu_r) \\ &< \frac{N_l}{N_l - H} d(\mathbf{X}_j, \mu_l) \end{aligned} \quad (8.12)$$

where  $N_r$  is the number of pixels in any cluster  $\omega_r$  and  $N_l$  that in cluster  $\omega_l$ .

If the clusters are assumed to have a normal distribution (Gaussian model), the Gaussian OMPR is formed as follows.

For all  $j$ , assign a data point  $\mathbf{X}_j$  in cluster  $\omega_l$  to cluster  $\omega_k$  if

$$\begin{aligned} \delta(\mathbf{X}_j, \omega_k) &= \min_{r \neq l} \delta(\mathbf{X}_j, \omega_r) \\ &< \ln|\Sigma_l| - \frac{N_l - H}{H} \ln \left[ 1 - \frac{H}{N_l - H} \Delta(\mathbf{X}_j, \omega_l) \right] \\ &- 2 \ln \frac{N_l}{N} - (D+2) \frac{N_l - H}{H} \ln \frac{N_l}{N_l - H} \end{aligned} \quad (8.13)$$

where

$$\delta(\mathbf{X}_j, \omega_r) = \ln|\Sigma_r| + \frac{N_r+H}{H} \ln \left[ 1 + \frac{H}{N_r+H} \Delta(\mathbf{X}_j, \omega_r) \right] - 2 \ln \frac{N_r}{N} + (D+2) \frac{N_r+H}{H} \ln \frac{N_r}{N_r+H} \quad (8.14)$$

and

$$\Delta(\mathbf{X}_j, \omega_r) = (\mathbf{X}_j - \mu_r)^T \Sigma_r^{-1} (\mathbf{X}_j - \mu_r)$$

with  $D$  the dimensionality of feature space  $\mathbf{X}$ .

In the OMPR method, data point inertia is considered. A data point shared by more pixels ('heavier') is more difficult to move from one cluster to another than a 'lighter' point.

## 8.5 Post-classification processing: smoothing and accuracy assessment

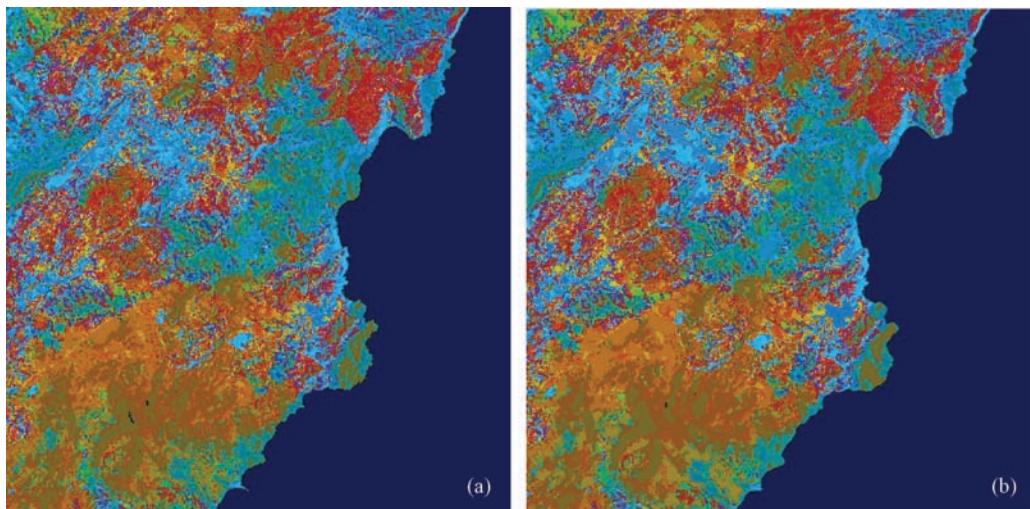
### 8.5.1 Class smoothing process

A classification image appears to be a digital image in which the DNs are the class numbers, but we cannot perform numerical operations on class numbers. For instance, the average of class 1 and class 2 cannot be class 1.5! Indeed, the class numbers in a

classification image do not have any sequential relationship; they are nominal values and can be treated as symbols such as  $A$ ,  $B$  and  $C$  (see also Section 12.3). A classification image is actually an image of symbols, *not* digital numbers; it is therefore *not* a digital image in the generally accepted sense. As such we cannot apply any numerical-operation-based image processing to classification images.

A classification image often contains noise caused by the isolated pixels of some classes, within another dominant class, which can form sizeable patches (Figure 8.7a). It is reasonable to presume that these isolated pixels are more likely to belong to this dominant class rather than to the classes that they are initially assigned to; these probably arise from classification errors. An appropriate smoothing process applied to a classification image will not only 'clean up' the image, making it visually less noisy, but also improve the accuracy of classification.

Among the low-pass filters that we have described so far, the only filter you can use to smooth a classification image is the *mode (majority) filter*. The reason for this is simple, since the mode filter smoothes an image without any numerical operations. For instance, if a pixel of class 5 is surrounded by pixels of class 2, the mode filter will reassign this pixel to class 2 according to the majority class in the



**Figure 8.7** Smoothing classification images: (a) ISODATA unsupervised classification with 24 classes; and (b) the classification image smoothed using a mode filter with a majority of 5 in a  $3 \times 3$  filtering kernel. A closer look at these images reveals the difference between them: image (b) is smoother than image (a)

filtering kernel. Figure 8.7b illustrates the effect of mode filtering applied to an unsupervised classification image in Figure 8.7a.

### 8.5.2 Classification accuracy assessment

Ultimately there is no satisfactory method to assess the absolute accuracy of image classification for remote sensing Earth observation applications (see also Section 17.5.1). The paradox is that we cannot conduct such an assessment without knowing 100% of ground truth on one hand, while, on the other hand, if we do have complete knowledge of ground truth, what is the point of the classification? Even an assessment or an estimate of relative accuracy of classification does, however, provide valuable knowledge for us to accept or reject a classification result at a certain confidence level. There are two generally accepted approaches to generate ground truth:

1. Use field-collected data of typical classes as samples of ground truth. For rapidly and temporally changing land cover classes, such as crops, field data should be collected simultaneously with image acquisition. For temporally stable targets, such as rocks and soils, published maps as well as field data can be used. The classification accuracy of the sample areas with known classes gives an estimate of total classification accuracy. This seemingly straightforward approach is often impractical in reality because it is often constrained by errors in the recording of

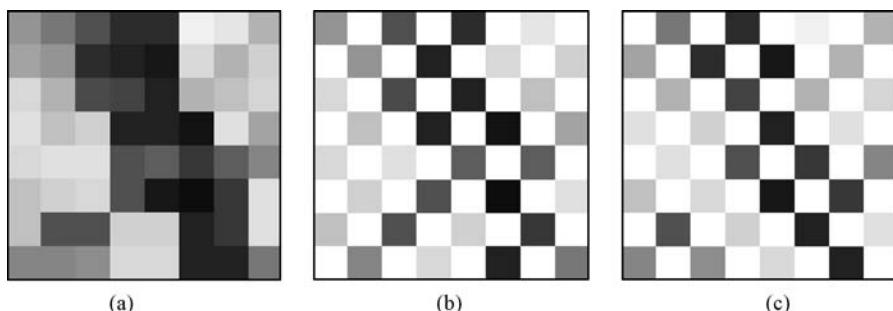
field observations, limited field accessibility and temporal irrelevance.

2. Another approach relies on image training. This uses typical spectral signatures and limited field experience, where a user can manually specify training areas of various classes using a multispectral image. The pixels in these training areas are separated into two sets: one is used to generate class statistics for supervised classification and the other for subsequent classification accuracy assessment. For a given training area, we could take a selection of pixels sampled from a  $2 \times 2$  grid as the training set while the remaining pixels are used for the verification set (ground truth reference data), as shown in Figure 8.8. The pixels in the verification set are assumed to belong to the same class as their corresponding training set. In another way, we can also select several training areas for the same class and use some of them for training and the rest for verification.

In practice the above two approaches are often used in combination.

Suppose that we have some kind of ground truth reference data; then a widely used method to describe the relative accuracy of classification is the *confusion matrix*:

$$\begin{pmatrix} C_{11} & C_{12} & \cdots & C_{im} \\ C_{21} & C_{22} & \cdots & C_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ C_{m1} & C_{m2} & \cdots & C_{mm} \end{pmatrix}. \quad (8.15)$$



**Figure 8.8** A resampling scheme for classification accuracy assessment. An image (a) is resampled to formulate two images (b) and (c); one is used as the training dataset while the other is used as the verification set

Here, each of the elements,  $C_{ii}$ , in the major diagonal represents the number of pixels that are correctly classified for class  $i$ . Any element off the major diagonal,  $C_{ij}$ , represents the number of pixels that should be in class  $i$  but which are incorrectly classified as class  $j$ . Obviously, if all the image pixels are correctly classified, we should then have a diagonal confusion matrix where all non-diagonal elements become zero. The sum of all the elements in the confusion matrix is the total number of pixels in a classification image,  $N$ :

$$N = \sum_{i=1}^m \sum_{j=1}^m C_{ij}.$$

The ratio between the summation of the major diagonal elements and the total number of pixels represents the percentage of the correct classification or *overall accuracy*:

$$\text{Ratio}_{\text{correct}}(\%) = \frac{1}{N} \sum_{i=1}^m C_{ii}. \quad (8.16)$$

The sum of any row  $i$  of the confusion matrix gives the total number of pixels that, according to the ground truth reference, should be in class  $i$ ,  $Nr_i$ :

$$Nr_i = \sum_{j=1}^m C_{ij}.$$

Then the ratio  $C_{ii}/Nr_i$  is the percentage of correct classification of class  $i$ , according to the ground truth references and is often called *user's accuracy*.

The sum of any column  $j$  of the confusion matrix gives the total number of pixels that have been classified as class  $j$ ,  $Nc_j$ :

$$Nc_j = \sum_{i=1}^m C_{ij}.$$

Then the ratio  $C_{ii}/Nc_j$  is the percentage of correct classification of class  $j$ , based on the classification result, and is often called *producer's accuracy*.

Apart from the above accuracy measurements which are based on simple ratios, another commonly used statistical measure of classification accuracy and quality is the *kappa coefficient* ( $\kappa$ ) that combines the above two class accuracy estimations, based on the rows and columns of the confusion

matrix, to produce an estimate of total classification accuracy, as follows:

$$\kappa = \frac{N \sum_{i=1}^m C_{ii} - \sum_{i=1}^m Nr_i \cdot Nc_i}{N^2 - \sum_{i=1}^m Nr_i \cdot Nc_i}. \quad (8.17)$$

In the case of 100% agreement between the classification and the reference data, the confusion matrix is diagonal, that is  $\sum_{i=1}^m C_{ii} = N$ . Thus,

$$\kappa = \frac{N^2 - \sum_{i=1}^m Nr_i \cdot Nc_i}{N^2 - \sum_{i=1}^m Nr_i \cdot Nc_i} = 1,$$

while if there is no agreement at all, then all the elements on the diagonal of the confusion matrix are zero, that is  $\sum_{i=1}^m C_{ii} = 0$ . In this case

$$\kappa = \frac{- \sum_{i=1}^m Nr_i \cdot Nc_i}{N^2 - \sum_{i=1}^m Nr_i \cdot Nc_i} < 0.$$

In summary, the maximum value of the kappa coefficient  $\kappa$  is 1, indicating perfect agreement between the classification and the reference data, while for no agreement  $\kappa$  becomes negative. The minimum value of  $\kappa$  is case dependent, but as long as  $\kappa \leq 0$ , it indicates zero agreement between the classification and the reference data.

As illustrated in Table 8.1, the numbers in bold italics form the confusion matrix.  $Nr_i$  and  $C_{ii}/Nr_i$  are listed in the two right-hand columns, while  $Nc_j$  and  $C_{ij}/Nc_j$  appear in the bottom two rows. The bold number in the bottom right corner is the total percentage of correct classification. The kappa coefficient can then be calculated from Table 8.1 by

$$\kappa = \frac{403 \times 308 - 33\,023}{162\,409 - 33\,023} = \frac{911\,01}{129\,386} = 0.704.$$

Despite the fact that the classification accuracy derived from the confusion matrix is very much a self-assessment and is by no means the true accuracy of classification, it does provide a useful measure of classification accuracy. The information in a confusion matrix is highly dependent on the quality of the training areas and field data. Well-selected training

**Table 8.1** An example confusion matrix

Class Reference	Class 1	Class 2	Class 3	Class 4	Class 5	Row sum $Nr_i$	$C_{ii}/Nr_i$ (%)
Reference 1	<b>56</b>	<b>9</b>	<b>5</b>	<b>2</b>	<b>8</b>	80	70.0
Reference 2	<b>10</b>	<b>70</b>	<b>7</b>	<b>3</b>	<b>5</b>	95	73.7
Reference 3	<b>0</b>	<b>3</b>	<b>57</b>	<b>10</b>	<b>6</b>	76	75.0
Reference 4	<b>0</b>	<b>6</b>	<b>0</b>	<b>79</b>	<b>4</b>	89	88.8
Reference 5	<b>8</b>	<b>4</b>	<b>3</b>	<b>2</b>	<b>46</b>	63	73.0
Column sum $Nc_j$	74	92	72	96	69	403	
$C_{jj}/Nc_j$ (%)	75.6	76.1	79.2	82.3	66.7		<b>76.4</b>

areas can improve both the classification accuracy and the credibility of accuracy assessment, whereas poorly selected training areas will yield low classification accuracy and unreliable accuracy assessment. Strictly speaking, this method only gives an estimate of the classification accuracy of the whole image.

## 8.6 Summary

In this chapter, we have introduced the most commonly used image classification approaches and algorithms. These methods are essentially multi-variable statistical classifications that achieve data partition in the multi-dimensional feature space of multi-layer image data, such as a multi-spectral remotely sensed image.

The iterative clustering method of unsupervised classification enables self-optimization of a local optimal representative of the natural clusters in the data. How well the clustering converges to a local optimal depends on the dissimilarity function and clustering mechanism employed, while the quality of the local optimal is mainly affected by the initial cluster centres (the seeds) from where the iteration starts. Thus a seed selection technique, locating the peaks of data distribution, is introduced. A method for cluster splitting, based on PC1, is also proposed to improve the clustering mechanism.

Though affected by the same factors, the accuracy of a supervised classification is largely controlled by the user's knowledge. High-quality user knowledge could lead to correct classification of known targets while poor user knowledge may mislead rather than help.

There are many methods of accuracy assessment, such as the well-known confusion matrix, but it is important to know the limitations of such methods that merely give a relative assessment rather than the true accuracy of classification.

Finally, we must recognize that a classification image is not a true digital image but a symbol image presented in numbers. We could apply numerical operations to a classification but the results do not really make any sense. We can, however, use logical operations to process classification images, such as smoothing a classification image using a mode (majority) filter because it does not involve any numerical operations.

## Questions

- 8.1 What is multi-variable statistical classification? Describe the major approaches for image classification.
- 8.2 What are the advantages and disadvantages of unsupervised classification? Describe the algorithm  $\alpha$  for iterative clustering.
- 8.3 Explain, using a diagram, the self-optimization mechanism of iterative clustering.
- 8.4 Describe the main steps of the 3D-FSIC algorithm with the aid of diagrams. What are the main advantages and limitations of feature space iterative clustering?
- 8.5 What are the two properties for the design of the automatic seed selection technique?
- 8.6 What is the problem with cluster splitting along the axis of the variable with the

- maximum standard deviation? What is a better approach?
- 8.7 Describe the general steps of supervised classification.
- 8.8 Explain the principle of spectral angle classification and its merits.
- 8.9 What is a confusion matrix? Based on the confusion matrix, give definitions for overall accuracy, user's accuracy and producer's accuracy.
- 8.10 Comment on the issue of accuracy assessment for image classification.



# 9

## Image Geometric Operations

Geometric operations include the shift, rotation and warping of images to a given shape or framework. In remote sensing applications, geometric operations are mainly used for the co-registration of images of the same scene acquired by different sensor systems or at different times or from different positions, and for rectifying an image to fit a particular coordinate system (geocoding). Image mosaic is a geometric operation that was commonly used in the early days of remote sensing image processing when computer power was inadequate for the massive demands of the geocoding process, but this is no longer the case. Once a set of adjacent images is accurately rectified to a map projection system, such as a UTM coordinate system (see Chapter 13 in Part Two for details) the images, though separate, are effectively in a mosaic.

### 9.1 Image geometric deformation

An image taken from any sensor system is a distortion of the real scene. There are many sources of error since, for instance, any optical sensor system is a distorted filtered imaging system. Such source errors in sensor systems are usually corrected in the sensor calibrations carried out by the manufacturers during hardware maintenance; they are beyond the scope of this chapter. Our main concerns lie on the user side of remote sensing applications, in the

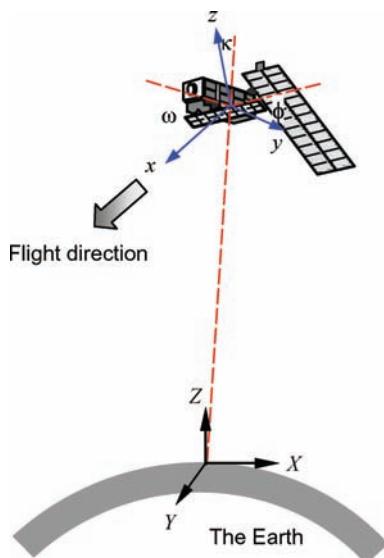
geometric distortions introduced during the imaging process, when a satellite or aircraft acquires images of the land surface.

#### 9.1.1 Platform flight coordinates, sensor status and imaging geometry

As shown in Figure 9.1, image geometry is fundamentally controlled by three sets of parameters:

- The platform flight coordinate system ( $x, y, z$ ), where  $x$  is in the flight direction,  $z$  is orthogonal to  $x$  in the plane through the  $x$  axis and perpendicular to the Earth's surface, and  $y$  is orthogonal to both  $x$  and  $z$ .
- The sensor 3D status is decided by orientation angles  $\omega, \phi, \kappa$  in relation to the platform flight coordinate system ( $x, y, z$ ).
- The coordinates ( $X, Y, Z$ ) of the imaging position on the ground usually conform to a standard coordinate system (defined by map projection and datum).

The preferred imaging geometry for the optical sensors of most Earth observation satellites is that the satellite travels horizontally and parallel to the Earth's curved surface with ( $x, y, z$ ) matching ( $X, Y, Z$ ) and sensor orientation angles ( $\omega, \phi, \kappa$ ) all being equal to 0. This is the configuration of nadir (vertical) imaging, which introduces minimal geometric distortion.

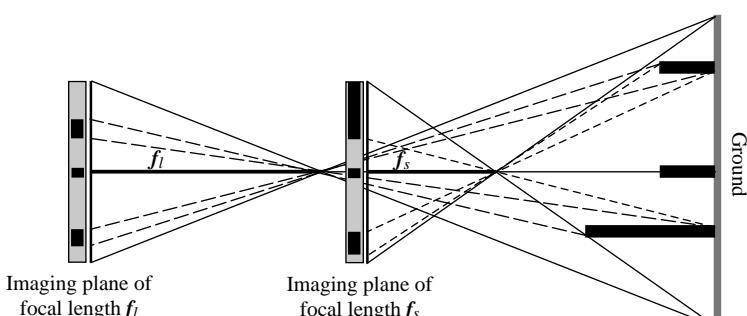


**Figure 9.1** Earth observation satellite imaging geometry. For the platform flight coordinate system  $(x, y, z)$ ,  $x$  is in the flight direction,  $z$  is orthogonal to  $x$  and in the plane through the  $x$  axis and perpendicular to the Earth's surface, and  $y$  is orthogonal to both  $x$  and  $z$ . The sensor status is decided by orientation angles  $(\omega, \phi, \kappa)$  in relation to the platform flight coordinate system  $(x, y, z)$ . The coordinates  $(X, Y, Z)$ , of the imaging position usually conform to a standard coordinate system (of defined map projection and datum)

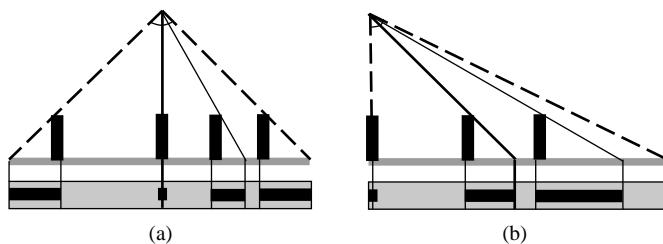
For an optical imaging system, the focal length  $f$  of the lenses is another important parameter that decides the characteristics of the central projection distortion. For the same imaging area (field of

view), a shorter focal length will result in greater topographic distortion (Figure 9.2). Nadir view imaging is achieved when the principal of the optical system is vertical to the plane of the scene to be imaged. Otherwise, an oblique view imaging configuration is formed, depending on the sensor status; this can be side-looking, forward-looking, backward-looking or oblique-looking in any direction. The geometric distortion for a nadir view image is central and symmetrical, increasing from the centre of the image to its edge, whereas that for an oblique view image increases from the near range to the far range of the image, in which all topographic features appear to fall away from the sensor look direction (Figure 9.3). For satellite-borne vertical imaging using a scanner system, the usual sensor status is that the orientation of the sensor principal is vertical to the land surface and that the scanning direction is perpendicular to the flight direction. As illustrated in Figure 9.4, when the scanning direction is skewed to the flight direction ( $\omega \neq 0$ ), the swath of the image scan line will be narrower than it is perpendicular to the flight direction. With this deliberate configuration, platform flight at a rotation angle from the flight direction can help to achieve higher spatial resolution at the cost of a narrower image scene, as in the case of SPOT 5 in high-resolution mode.

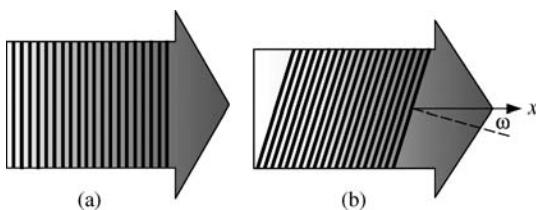
The sensor status  $(\omega, \phi, \kappa)$  parameters are the most sensitive to the image geometry. A tiny displacement of sensor status can translate to significant distortion in image pixel position and scale. Great technical effort has been devoted to achieve very precise status control for modern Earth



**Figure 9.2** The relationship between focal length and geometric distortion. A sensor with a shorter focal length ( $f_s$ ) can cover the same field of view in a much shorter distance than that with a longer focal length ( $f_l$ ) but with more significant geometric distortion of tall ground objects and high terrain relief



**Figure 9.3** Central projection distortion for nadir view (a) and oblique view (b) imaging



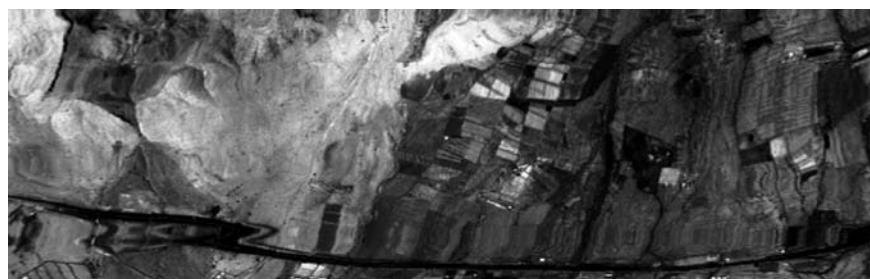
**Figure 9.4** The effects of sensor status in relation to image scanner orientation: (a) the scanning direction is perpendicular to the flight direction; and (b) the scanning direction is oblique to the flight direction, because the sensor/platform has a rotation angle relating to the flight direction. Consequently, the swath of the image becomes narrower

observation satellites, to reduce sensor status distortion to a minimum and thus to satisfy most applications. For airborne sensor systems, the stability of the platform is often problematic, as demonstrated in the airborne thematic mapper image in Figure 9.5. This type of error can be corrected using onboard flight status parameters, to some extent, though the errors are often too severe to be corrected for.

### 9.1.2 Earth rotation and curvature

For spaceborne Earth observation remote sensing, the sensor system onboard a satellite images a 3D spherical land surface with topographic relief onto a 2D flat image. Geometric distortion and position inaccuracy are inevitable. Fortunately, over centuries, photogrammetric (or geomatic) engineering has developed many effective map projection models to achieve the optimized translation of the Earth's spherical surface to a flat surface of maps and images. The most widely used map projection system is the UTM (Universal Transverse Mercator) with either the WGS84 global datum or a local datum. We are going to revisit the topic of map projection in greater detail in Chapter 13 in Part Two of the book. One of the major tasks of image geometric operation is to rectify an image to a given map projection system. The process is often called *geocoding* or *georeferencing*.

Many Earth observation satellites are configured to fly in circular, near-polar, Sun-synchronous orbits, so as to image nearly every part of the Earth at about the same local time. In such an orbit, a



**Figure 9.5** Image scan line distortion in an ATM image, as caused by aircraft yaw

satellite travels nearly perpendicular to the Earth's rotation direction. This is not a problem for an instant image taken by a camera as all the pixels in the scene are taken simultaneously. It becomes a problem for a scanner, however, which is still (so far) the dominant design for spaceborne sensor systems; the image is built line by line in a time sequence with the satellite flying above the rotating Earth. As shown in Figure 9.6, for an along-track push-broom scanner, the Earth's surface is imaged in consecutive swaths in the time interval of a scanning cycle (or scanning frequency) to build a sequence of image lines in the flight direction of the satellite. The Earth's rotation causes the longitude position to move westwards in each scanning cycle and, as a result, the image built up is not a rectangular stripe but a parallelogram stripe which is skewed to the west in the imaging advancing direction. For the scenario of an across-track two-way scanner (e.g. thematic mapper), the distortion pattern is more complicated because every pixel along a swath is imaged at a different time and the actual scanning speed at the Earth's surface changes not only from nadir to the edge of a swath but also between swaths for and against the Earth's rotation. The processing to compensate for the Earth's rotation in relation to the scanning mechanism is usually done in bulk geometric correction by the ground assembly facilities, e.g. the receiving station, using batch processing algorithms based on sensor/platform configuration parameters. The images after bulk processing to de-skew and crudely geocode are usually labelled as L-1B data.

## 9.2 Polynomial deformation model and image warping co-registration

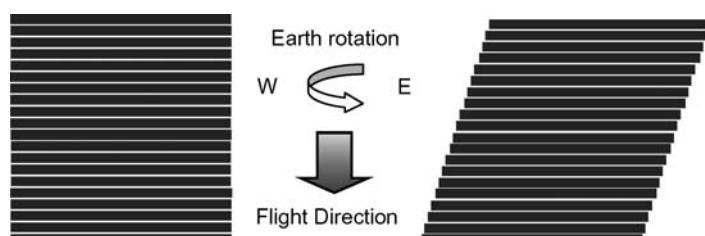
Based on the above discussion, we can generally assume that a remotely sensed image has some geometric distortion as a result of image acquisition. Given a reference map projection or a reference image which is either geometrically correct or regarded as a geometric basis of a set of images, the main task of geometric operations is to establish a deformation model between the *input image* and the *reference* and then rectify or co-register the input image to the reference to generate an *output image*. In this way, the geometric distortion of the input image is either corrected or co-registered to a standard basis for further analysis.

The so-called *rubber sheet warping*, based on a polynomial deformation model, is the most important and commonly used geometric transformation for remotely sensed image data. There are several types of image warping:

- Image to map projection system (e.g. UTM)
- Image to map (e.g. a topographic map)
- Image to image.

The geometric transformation includes two major steps:

1. Establish the polynomial deformation model. This is usually done using *ground control points* (GCPs).



**Figure 9.6** The skew effect of the Earth's rotation on a push-broom scanner onboard an Earth observation satellite in a circular, near-polar, Sun-synchronous orbit. The image is recorded as left but the actual area covered on the surface of the Earth is as right

2. Image resampling based on the deformation model. This includes resampling image pixel positions (coordinates) and DNs.

### 9.2.1 Derivation of deformation model

A deformation model can be derived by fitting a polynomial function to ground control point locations. This is done by selecting many GCPs representing the same ground positions in both the input image and the reference (an image or a map) to establish a deformation model, and then transforming the input image to the output image that is forced to fit the reference geometry.

In simple terms, this means transforming each pixel of the input image to the output image (input-to-output mapping) based on the deformation model, but a pixel position in an image is given in integers of line and column, while the transformation between the input and the output images may not always correspond exactly to integer positions for every pixel. Many pixels in the input image may take decimal positions, overlap or be apart from each other resulting in ‘black holes’ in the output image. The input-to-output mapping cannot therefore generate an output image as a proper, regular raster image. To solve the problem, a commonly used approach is *output-to-input mapping*.

Suppose transformation  $M$  is an output-to-input mapping that maps an output pixel at position  $(i, j)$  back to the input image at location  $(k, l)$ ; then the output image can be generated as shown in Figure 9.7. For each output pixel position  $(i, j)$ , computation from  $M(i, j)$  gives the corresponding position  $(k, l)$  in the input image and then, at this position, a pixel value  $P_i(k, l)$  is picked up and assigned to the output pixel  $P_o(i, j)$ . The output image is completed when all the pixels have been assigned.

The question now is how to derive a deformation model, or the transformation  $M$ . Let  $(k, l)$  represent the position in the input image corresponding to the output position  $(i, j)$ ; then the general form of the polynomial approximation for  $k$  and  $l$  is

$$M : \begin{cases} k = Q(i, j) \\ \quad = q_0 + q_1 i + q_2 j + q_3 i^2 + q_4 i j + q_5 j^2 + \dots \\ l = R(i, j) \\ \quad = r_0 + r_1 i + r_2 j + r_3 i^2 + r_4 i j + r_5 j^2 + \dots \end{cases} \quad (9.1)$$

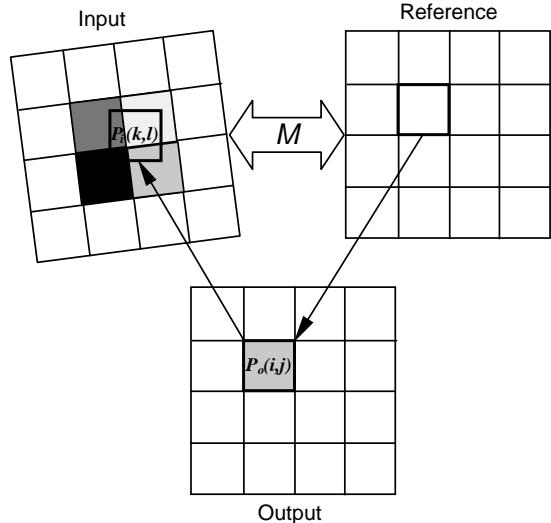


Figure 9.7 Output-to-input mapping

Formula (9.1) defines the transformation  $M$  that calculates the approximation of the input position  $(k, l)$  from a given output position  $(i, j)$ , if the coefficients  $\mathbf{Q} = (q_0, q_1, q_2, \dots)^T$  and  $\mathbf{R} = (r_0, r_1, r_2, \dots)^T$  are known. For  $n$  pairs of GCPs, we already know both  $(k, l)$  and  $(i, j)$  for every GCP, so the least squares solutions for  $\mathbf{Q}$  and  $\mathbf{R}$  can be derived. From  $n$  GCPs, we can establish  $n$  pairs of polynomials based on (9.1) written in matrix format as

$$\begin{aligned} \mathbf{K} &= \mathbf{MQ} \\ \mathbf{L} &= \mathbf{MR} \end{aligned} \quad (9.2)$$

where

$$\mathbf{K} = \begin{pmatrix} k_1 \\ k_2 \\ \vdots \\ k_n \end{pmatrix}, \quad \mathbf{L} = \begin{pmatrix} l_1 \\ l_2 \\ \vdots \\ l_n \end{pmatrix}, \quad \mathbf{Q} = \begin{pmatrix} q_0 \\ q_1 \\ q_2 \\ \vdots \end{pmatrix},$$

$$\mathbf{R} = \begin{pmatrix} r_1 \\ r_2 \\ r_3 \\ \vdots \end{pmatrix}, \quad \mathbf{M} = \begin{pmatrix} 1 & i_1 & j_1 & i_1^2 & j_1^2 & \dots \\ 1 & i_2 & j_2 & i_2^2 & j_2^2 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & i_n & j_n & i_n^2 & j_n^2 & \dots \end{pmatrix}.$$

The least squares solution for  $\mathbf{Q}$  is

$$\mathbf{Q} = (\mathbf{M}^T \mathbf{M})^{-1} \mathbf{M}^T \mathbf{K}. \quad (9.3)$$

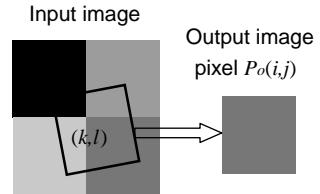
Similarly, for  $\mathbf{R}$ ,

$$\mathbf{R} = (\mathbf{M}^T \mathbf{M})^{-1} \mathbf{M}^T \mathbf{L}. \quad (9.4)$$

Once the coefficients  $\mathbf{Q} = (q_0, q_1, q_2, \dots)^T$  and  $\mathbf{R} = (r_0, r_1, r_2, \dots)^T$  are derived from the GCPs, the pixel position relationship between the input and output images is fully established through the transformation  $M$ . Given a location  $(i, j)$  in the output image, the corresponding position  $(k, l)$  in the input image can then be calculated from the transform  $M$  using Equation (9.1). Theoretically, the higher the order of the polynomials, the higher the accuracy of warping that can be achieved, but the more the number of control points needed. A linear fit needs at least three GCPs, a quadric fitting six and a cubic fitting ten.

### 9.2.2 Pixel DN resampling

In the output-to-input mapping model, the output pixel at position  $(i, j)$  is mapped to its corresponding position  $(k, l)$  in the input image by the transform  $M$ . In most cases,  $(k, l)$  is not an integer position and there is no pixel DN value ready for this point. Resampling is an interpolation procedure used to find the DN for position  $(k, l)$  in the input image so as to assign it to the pixel  $P_o(i, j)$  in the output image



**Figure 9.8** Illustration of nearest neighbour DN resampling. The nearest neighbour of the output pixel  $P_o(i, j)$  in the input image is the pixel at the bottom right and therefore the DN of this pixel is assigned to  $P_o(i, j)$

(Figure 9.7). The simplest resampling function is the *nearest neighbour method* (Figure 9.8), in which pixel  $P_o(i, j)$  is assigned the DN of the input image pixel nearest to position  $(k, l)$ .

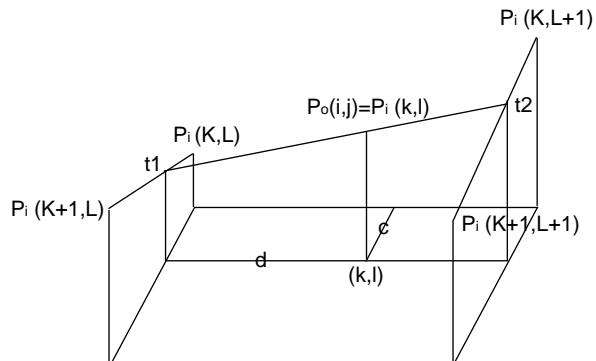
A more accurate and widely used method is *bilinear interpolation*, as defined below (Figure 9.9):

$$\begin{aligned} t_1 &= P_i(K, L)(1 - c) + P_i(K + 1, L)c \\ t_2 &= P_i(K, L + 1)(1 - c) + P_i(K + 1, L + 1)c \\ P_o(i, j) &= t_1(1 - d) + t_2d \end{aligned} \quad (9.5)$$

where,

$P_i(K, L)$  is an input pixel at an integer position  $(K, L)$  in the input image

$P_o(i, j)$  is the output pixel at an integer position  $(i, j)$  in the output image, which corresponds to  $P_i(k, l)$  at a decimal position  $(k, l)$  in the input image



**Figure 9.9** Illustration of bilinear interpolation

$K$  is the integer part of  $k$

$L$  is the integer part of  $l$

$$c = k - K$$

$$d = l - L.$$

Quadratic and cubic polynomials are also popular interpolation functions for resampling with more complexity and improved accuracy.

## 9.3 GCP selection and automation

### 9.3.1 Manual and semi-automatic GCP selection

GCPs can be selected manually, semi-automatically or automatically. These points are typically distinctive corner points with sharp contrast to their surroundings. For the manual method, the accuracy of the GCPs depends totally on the user's shaking hands. It is often not easy to locate accurately the corresponding points in the input and reference images. A simple approach can improve the efficiency and accuracy for manual GCP selection. Firstly, select four GCPs spread in the four corners of the input and the reference images and thus an initial geocoding frame (a linear fitting transformation  $M$ ) is set up based on these four GCPs. Using four instead of the minimum three GCPs for the linear fitting can allow initial error estimation. After this initial setup, once a GCP is selected in the reference image, the corresponding position in the input image is roughly located based on the initial geocoding frame. A user only needs to fine-tune the position and confirm the selection. The transformation  $M$  will be continuously updated as GCPs are added, via the least squares solution.

The semi-automatic method allows users to identify GCPs in corresponding (input and reference) images and then automatically optimize point positions from one image to the other using local correlation. Suppose the GCP in the input image is at a position  $(k, l)$ ; the optimal coordinates of this point in the reference image are then decided when  $r(k, l)$ , the normalized cross-correlation (NCC) coefficient between the input image and reference image at position  $(k, l)$ , reaches the maximum in an  $l_w \times s_w$  calculation window used to roam in an  $l_s \times s_s$  searching area

in the reference image surrounding the roughly selected position of the GCP:

$$r(k, l) = \frac{\sum_{i=1}^{l_w} \sum_{j=1}^{s_w} (w_{i,j} - \bar{w})(s_{k-1+i, l-1+j} - \bar{s}_{k,l})}{\left\{ \begin{array}{l} \left[ \sum_{i=1}^{l_w} \sum_{j=1}^{s_w} (w_{i,j} - \bar{w})^2 \right]^{1/2} \\ \times \left[ \sum_{i=1}^{l_w} \sum_{j=1}^{s_w} (s_{k-1+i, l-1+j} - \bar{s}_{k,l})^2 \right]^{1/2} \end{array} \right\}} \Rightarrow \max \quad (9.6)$$

where  $(w_{i,j} - \bar{w})$  is calculated from the input image while  $(s_{k-1+i, l-1+j} - \bar{s}_{k,l})$  is from the reference image.

Instead of NCC, a relatively new technique, phase correlation, can be used to locate directly the matching pixels in corresponding images at sub-pixel accuracy without roaming search. A totally different image co-registration method based on phase-correlation-derived sub-pixel optical flow will be introduced in Section 9.4.

### 9.3.2 \*Towards automatic GCP selection

Automatic GCP selection is an ongoing topic of research. It belongs to a very active research field in computer vision: feature tracking. Automatic GCP selection enables the automation of image-to-image co-registration. A workable automatic GCP selection method must be able to accurately select adequate high-quality GCPs that are evenly spread across the whole image to be co-registered. In general, it comprises two steps:

1. Automatic selection of corner points from one of the two corresponding images (either the input image or the reference image) as candidate GCPs.
2. Automatic location of the corresponding points in the other image via local correlation. High-quality GCPs are finally selected by their correlation level and spatial distribution.

The first step is the key to the automation of GCP selection. When manually selecting GCPs, we typically look for corner points such as corners of buildings, road junctions, sharp river bends and



**Figure 9.10** Five types of corner points: L-corner, Y-junctions, T-junctions, arrow-junctions and X-junctions

sharp topographic features. These points in images are of high contrast and can be characterized into five types: L-corner, Y-junctions, T-junctions, arrow-junctions and X-junctions, as presented in Figure 9.10.

Great effort has been put into developing effective techniques for corner point detection from digital images. Some of the more commonly used techniques are based on an autocorrelation matrix. Following the convention in the description of autocorrelation, we use  $I(x, y)$  to denote an image function (the DN of any an image pixel at position  $x, y$ ). Given a shift  $(\Delta x, \Delta y)$ , the autocorrelation at pixel  $(x, y)$  is defined as

$$c(x, y) = \sum_W [I(x_i, y_i) - I(x_i + \Delta x, y_i + \Delta y)]^2 \quad (9.7)$$

where  $I(x_i, y_i)$  denotes the image grey value at pixel position  $(x_i, y_i)$  in the 2D calculation window  $W$  (with a Gaussian weighting function) centred at  $(x, y)$ .

Consider a first-order approximation based on the Taylor expansion:

$$I(x_i + \Delta x, y_i + \Delta y) \approx I(x_i, y_i)$$

$$+ [I_x(x_i, y_i)I_y(x_i, y_i)] \begin{pmatrix} \Delta x \\ \Delta y \end{pmatrix}. \quad (9.8)$$

Combining (9.7) and (9.8) yields

$$\begin{aligned} c(x, y) &= \sum_W \left[ [I_x(x_i, y_i)I_y(x_i, y_i)] \begin{pmatrix} \Delta x \\ \Delta y \end{pmatrix} \right]^2 \\ &= (\Delta x \Delta y) \begin{pmatrix} \sum_W (I_x(x_i, y_i))^2 & \sum_W I_x(x_i, y_i)I_y(x_i, y_i) \\ \sum_W I_x(x_i, y_i)I_y(x_i, y_i) & \sum_W (I_y(x_i, y_i))^2 \end{pmatrix} \begin{pmatrix} \Delta x \\ \Delta y \end{pmatrix} \\ &= (\Delta x \Delta y)A(x, y) \begin{pmatrix} \Delta x \\ \Delta y \end{pmatrix}. \end{aligned} \quad (9.9)$$

Equation (9.9) shows that the autocorrelation function can be approximated by the autocorrelation matrix  $A(x, y)$ . This matrix captures the structure of the neighbourhood, and its eigenvalues  $\lambda_1$  and  $\lambda_2$  are proportional to the principal curvatures of the local autocorrelation function and form a rotation-invariant description of the matrix  $A(x, y)$ . Based on the magnitudes of the eigenvalues, the following inferences can be made:

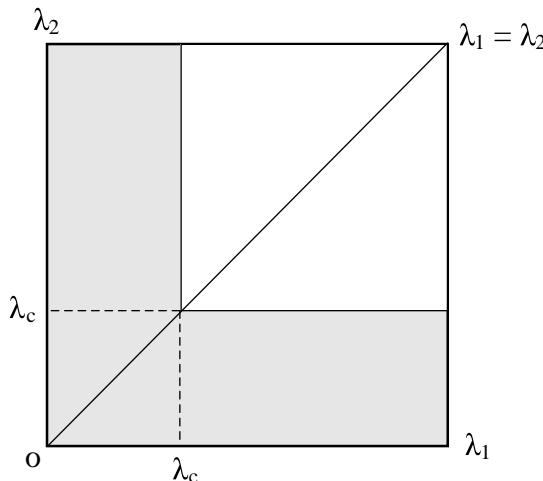
1. If  $\lambda_1 = 0$  and  $\lambda_2 = 0$ , the matrix is of rank 0; there are no features of interest at the image position  $(x, y)$  that is in a flat region.
2. If  $\lambda_1 = 0$  and  $\lambda_2$  is some large positive value, the matrix is of rank 1 and an edge is found where no changes occur along the edge direction.
3. If  $\lambda_1$  and  $\lambda_2$  are both large, distinct, positive values, the matrix is a full-rank matrix. Significant changes are found in orthogonal directions and a corner is found.

Based on the above properties of the autocorrelation matrix  $A(x, y)$ , many corner detectors were proposed and, among them, a simple and effective technique based on Shi and Tomasi (1994) is introduced as

$$\min(\lambda_1, \lambda_2) > \lambda_c \quad (9.10)$$

where  $\lambda_c$  is a predefined criterion of the minimum eigenvalue.

An image pixel is accepted as a good corner point if its autocorrelation matrix satisfies the condition set by formula (9.10) that assures that both eigenvalues are sufficiently large. Figure 9.11 illustrates the acceptance area in  $(\lambda_1, \lambda_2)$  coordinates for a given  $\lambda_c$ . The range of  $\lambda_1$  and  $\lambda_2$  is usually  $[0, 15\,000]$  for typical remotely sensed images. Once adequate high-quality corner points are



**Figure 9.11** For a given criterion  $\lambda_c$ , the corner detector defined in formula (9.10) accepts the points in the white area and rejects those in the shaded area

selected from one image, then the matching points in the corresponding image for co-registration can be located by local correlation through either NCC or phase correlation. As both the automatic and semi-automatic GCP selection methods must be based on local correlation, they are not suitable for co-registration between an image and a digital map as map features are synoptic and not necessarily correlated to image features. Methods of this type perform poorly where there are significant differences in scale and spatial resolution between the images being co-registered.

## 9.4 \*Optical flow image co-registration to sub-pixel accuracy

The classical approach to image warping co-registration, based on a polynomial deformation model derived from GCPs, can be quite accurate if the GCPs are of high positional accuracy, but registration by this method is not achieved at the pixel-to-pixel level. The registration error within a scene may vary from place to place according to the local relative deformation between the images. The imaging-geometry-related systematic deformation can be modelled and compensated for, but the

irregular deformation caused by sensor optical quality, platform status precision and so on cannot be effectively corrected using GCP deformation-model-based approaches.

A different approach for image co-registration is the forced pixel-to-pixel image co-registration method (Lucas and Kanade, 1981). Instead of using a transform grid determined by the GCP coordinate transformation, this approach achieves image co-registration pixel by pixel via image local spatial feature matching. This is fundamentally different from image warping and constitutes image co-registration but not georectification. In other words, an image cannot be transformed to a given map projection based on the map coordinates using this method, whereas it can be done using GCP-based techniques. The pixel-to-pixel image co-registration can achieve image georectification only when the reference image is geocoded. The advantage of precise feature matching at the pixel-to-sub-pixel level makes the technique a superior tool for very demanding quantitative change detection tasks, such as monitoring the spatial, geometric, positional and spectral changes of sensitive ground targets from space.

Precise image feature matching has to be done at a sub-pixel level as it is unlikely that the feature mismatching between two images will always be at an integer pixel position; as indicated earlier, a pixel in one image often corresponds to a decimal pixel position in the other. Techniques that can achieve image feature matching at sub-pixel accuracy hold the key for pixel-to-pixel image co-registration.

The recent development of phase correlation algorithms, namely singular value decomposition (SVD) (Hoge, 2003) and robust 2D fitting (Balci and Foroosh, 2005; Liu and Yan, 2006), enables calculation of image shift, rotation and scale change at better than 1/20th pixel accuracy. Based on these algorithms, a robust pixel-to-pixel image co-registration method is developed (Liu and Yan, 2008).

### 9.4.1 Basics of phase correlation

Phase correlation is based on the well-known Fourier shift property: a shift in the spatial coordinate frame between two functions results in a linear

phase difference in the frequency domain of the Fourier transform (FT) of the two functions. Given two 2D functions  $g(x, y)$  and  $h(x, y)$  representing two images related by a simple translational shift  $a$  in the horizontal and  $b$  in the vertical direction,  $h(x, y) = g(x + a, y + b)$ , and the corresponding FTs are denoted  $G(u, v)$  and  $H(u, v)$ , then

$$H(u, v) = G(u, v)e^{-i(au + bv)}. \quad (9.11)$$

The phase correlation is defined as the normalized cross-power spectrum between  $G$  and  $H$ , which is a matrix:

$$Q(u, v) = \frac{G(u, v)H(u, v)^*}{|G(u, v)H(u, v)^*|} = e^{-i(au + bv)}. \quad (9.12)$$

If  $G(u, v)$  and  $H(u, v)$  are continuous functions, then the inverse Fourier Transform (IFT) of  $Q(u, v)$  is a delta function:

$$q(x, y) = \delta(x - a, y - b) \quad (9.13)$$

where the function peak identifies the magnitude of the shift (Stiller and Konrad, 1999). For the case of raster data of imagery, which is a discrete 2D function,  $q(x, y)$  presents a delta-like function as if  $a$  and  $b$  are integers and, subsequently, the translation estimate between the two similar images can only be performed at integer (pixel) accuracy even though the true shifts  $a$  and  $b$  may well be real numbers with decimal parts (or sub-pixels). The delta function defined by (9.13) is therefore not an ideal solution.

As the magnitude of  $Q(u, v)$  is normalized to 1, the only variable in (9.12) is the phase shift defined by  $au + bv$  where  $a$  and  $b$  are the horizontal and vertical magnitudes of the image shift between  $g(x, y)$  and  $h(x, y)$ . Obviously, if we can solve  $a$  and  $b$  accurately based on the phase correlation matrix  $Q(x, y)$ , then the non-integer translation estimate at sub-pixel accuracy can be achieved in the frequency domain without the IFT. Two algorithms are proposed: the SVD method and the 2D fitting method. Such a direct frequency domain solution has proven more accurate and faster than that based on the delta function of (9.13) via IFT. The SVD method deals better with the large magnitude of image frame shifts and low correlation, while the robust 2D fitting method is much faster.

For image rotation, consider image  $f_2$  as a replica of image  $f_1$  with rotation  $\theta_0$ . Then they are related by

$$f_2(x, y) = f_1(x \cos\theta_0 + y \sin\theta_0, -x \sin\theta_0 + y \cos\theta_0). \quad (9.14)$$

The Fourier rotation property shows that the FTs between  $f_1$  and  $f_2$  are related by

$$F_2(\xi, \eta) = F_1(\xi \cos\theta_0 + \eta \sin\theta_0, -\xi \sin\theta_0 + \eta \cos\theta_0). \quad (9.15)$$

If the frequency domain is represented in polar coordinates, then the rotation will be an angular shift on the angle coordinate. In a polar coordinates system, we then have

$$F_2(\rho, \theta) = F_1(\rho, \theta - \theta_0). \quad (9.16)$$

The rotation can thus be found as a phase shift in the frequency domain that again can be determined by phase correlation (Reddy and Chatterji, 1996). As rotation involves significant pixel shift away from the rotation centre, only the SVD algorithm is robust enough to measure it (Liu and Yan, 2006).

If image  $f_2$  is a replica of  $f_1$  scaled by  $(a, b)$ , then they are related by

$$f_2(x, y) = f_1(ax, by). \quad (9.17)$$

The Fourier scale property shows the transforms are related by

$$F_2(\xi, \eta) = \frac{1}{|ab|} F_1\left(\frac{\xi}{a}, \frac{\eta}{b}\right). \quad (9.18)$$

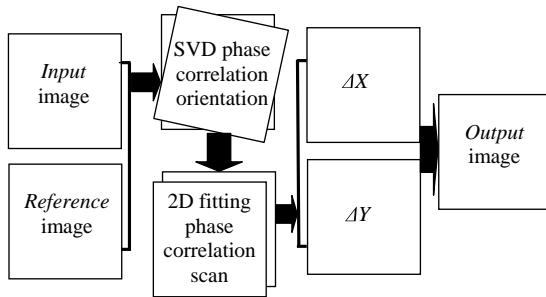
Ignoring the multiplicative factor and taking logarithms,

$$F_2(\log \xi, \log \eta) = F_1(\log \xi - \log a, \log \eta - \log b). \quad (9.19)$$

Thus a change in scale can be determined based on a phase shift in the frequency domain presented in logarithmic coordinate units and thus again can be determined by phase correlation (Reddy and Chatterji, 1996).

#### 9.4.2 Basic scheme of pixel-to-pixel image co-registration

The SVD algorithm is robust for the orientation and matching of a pair of large images with considerable



**Figure 9.12** Schematic illustration of phase correlation pixel-to-pixel co-registration

frame shift, rotation and minor scale change, and relatively low spectral correlation, while the 2D fitting is much faster and more reliable in small-window feature matching. With these considerations in mind, a phase correlation pixel-to-pixel image co-registration scheme is outlined below and illustrated in Figure 9.12:

- Given two images for co-registration, set one as the *Input* and the other the *Reference*, and then the *Output* image frame is set based on the *Reference* image. If the pixel sizes of the two images are different, the lower resolution image should be oversampled to the same pixel size of the higher resolution image.
- Apply the SVD phase correlation algorithm to a square region roughly covering the common area between the *Input* and the *Reference* images. This will determine the frame shift, rotation and scale change between these two images. Then the *Input* image is roughly oriented to the *Reference* image by standard routines of image shift, rotation and scale change. The image rotation and minor scale change may introduce geometric errors from interpolation and resampling but these will largely be eliminated by the later steps of the pixel-to-pixel image co-registration based on the precise measurements of the shift between every pair of corresponding pixels.
- Scan the oriented pair of images using 2D fitting phase correlation with a  $64 \times 64$  calculation window, from the first effective calculation position (the pixel at line 32 and column 32 of the upper left corner of the overlapped image area) to the last effective calculation position (the pixel that is at line 32 and column 32 to the

bottom-right corner of the overlapped image area). Thus the relative column and line shifts ( $\delta_x, \delta_y$ ) between any pair of corresponding pixels in the *Input* and the *Reference* images are determined. This process produces the optical flow images of local feature shift in column and line directions,  $\Delta X$  and  $\Delta Y$ .

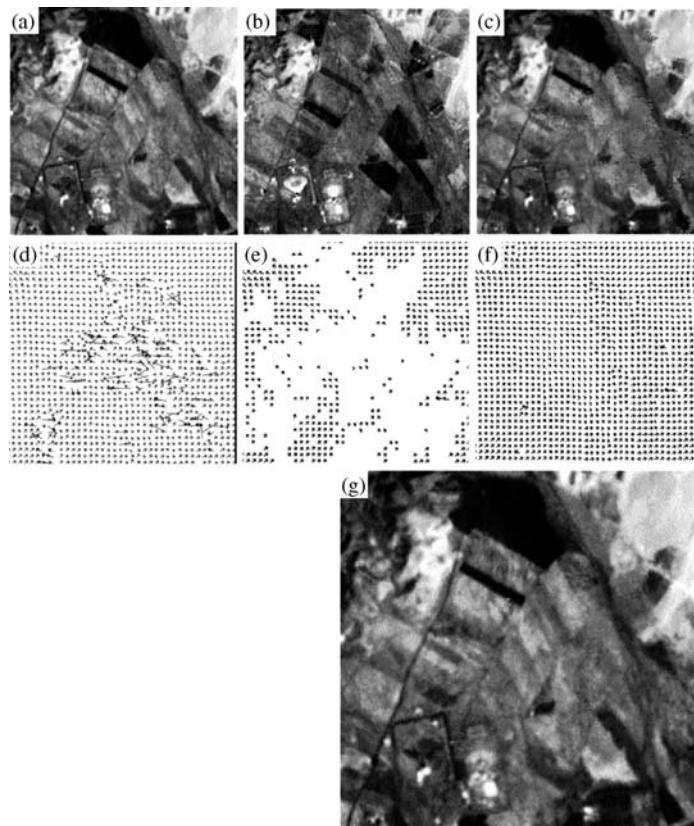
- Then an *Output-to-Input* mapping procedure is employed to carry out pixel-to-pixel co-registration, rectifying the *Input* image precisely to match the *Reference* image. This process builds up an *Output* image based on the *Reference* pixel positions and *Input* pixel DNs. Specifically, at each integer pixel position of the *Output* image, the pixel DN is drawn from its precise corresponding position in the *Input* image based on the shifts ( $\delta_x, \delta_y$ ), where  $\delta_x \in \Delta X, \delta_y \in \Delta Y$ , via bilinear interpolation.

The above scheme registers the *Input* image to the *Reference* image pixel-by-pixel to formulate the *Output* (co-registered) image. With the sub-pixel resolution of  $(\delta_x, \delta_y)$  calculated via phase correlation, the *Output* image matches the *Reference* image at a sub-pixel resolution at every pixel.

#### 9.4.3 The median shift propagation technique

As shown in Figure 9.13, the phase correlation may fail in featureless areas and areas of significant spectral differences between the *Input* image (Figure 9.13a) and the *Reference* image (Figure 9.13b) because of low correlation. As a result, drastic errors of registration may occur in the *Output* image (Figure 9.13c) because the feature shifts in these areas, measured using phase correlation, are incorrect. These errors are illustrated by feature shift vectors, of random direction and magnitude, in the optical flow  $\Delta X$  and  $\Delta Y$  (Figure 9.13d). Image pixels in these areas are characterized by their low correlations.

As the magnitude of phase correlation is unified, it does not give a direct measure of correlation quality. The correlation quality can, however, be reliably assessed based on the 2D fitting quality via the regression coefficient of determination (RCD) in a normalized value range [0, 1]. A correlation quality assessment image can then be generated together with the shift images  $\Delta X$  and  $\Delta Y$  by 2D fitting phase correlation scanning. Thus the



**Figure 9.13** Illustration of the MSP technique: (a) the Input image, ETM+ Pan; and (b) the Reference image, ASTER band 3. The two images have significant spectral differences and several featureless areas. (c) The initial *Output* image is contaminated by mis-registration errors in the featureless and spectrally different areas. (d) The vector optical flow of  $\Delta X$  and  $\Delta Y$  (resampled at 8 pixel interval and magnitude enlarged by four times for illustration) indicates that these errors result from the malfunction of phase correlation in feature shift measurement. (e) The low-correlation quality areas are masked off based on  $RCD < 0.75$ . (f) The gaps are smoothly filled using the MSP. (g) The final *Output* image shows that the co-registration errors are effectively eliminated

low-quality shift measurement data can be identified and masked off with a *Null* value from  $\Delta X$  and  $\Delta Y$  by an RCD threshold, say at  $RCD < 0.75$  (Figure 9.13e).

Now the problem is how to fill the gaps in the masked  $\Delta X$  and  $\Delta Y$  images with shift values of reasonable quality. A simple and effective solution for the problem is the median shift propagation (MSP) as detailed below (Liu and Yan, 2008):

1. Scan the masked  $\Delta X$  and  $\Delta Y$  images to search for a gap.
2. Whenever a *Null* value is encountered, a median filtering mechanism in a calculation window, for example  $5 \times 5$  pixels, is applied to the shift

images  $\Delta X$  and  $\Delta Y$ . This process replaces the *Null* with the median of the values of non-*Null* value pixels in the calculation window.

3. Continue the median filtering on the modified  $\Delta X$  and  $\Delta Y$ , till all the gaps are filled, to form a smooth optical flow in Figure 9.13f. The key difference between the MSP and ordinary median filtering is that instead of always applying the filter to the original  $\Delta X$  and  $\Delta Y$  image, the filter is applied to the  $\Delta X$  and  $\Delta Y$  images that are modified by the last filtering action. In such a way, the  $\Delta X$  and  $\Delta Y$  images are updated continuously during the filtering process and the feature shifts are thus self-propagated from high-quality data to fill the *Null* value gaps.

1.63	1.60	1.62	1.61	1.59	1.57	1.58	1.63	1.60	1.62	1.61	1.59	1.57	1.58
1.59	1.57	1.55	1.52	1.53	1.51	1.52	1.59	1.57	1.55	1.52	1.53	1.51	1.52
1.61	1.58						1.61	1.58	1.58	1.57	1.57		
1.56	1.54						1.56	1.54	1.56	1.55	1.55		
1.55	1.53						1.55	1.53	1.55	1.55	1.56		
1.52	1.51						1.52	1.51					
1.50	1.49						1.50	1.49					

**Figure 9.14** Numerical explanation of the MSP: (a) the NCC masked shift image (e.g.  $\Delta X$ ) with good-quality data in bold font and gaps; and (b) the numbers in italics in the central  $3 \times 3$  box with a bold frame are filled via the MSP from the existing quality data.

- With the  $\Delta X$  and  $\Delta Y$  images refined by the MSP, the *Input* image patches in the low-correlation areas are correctly rectified, by the geometry of the *Reference* image, into the *Output* image (Figure 9.13g).

Though we borrowed the term ‘median filter’, the process is not filtering but self-propagating via the median. As shown in Figure 9.14a, the  $3 \times 3$  gap area in the box with a bold frame can be filled gradually via median propagation from the data in the top two lines and left two columns using a  $5 \times 5$  processing window. The first cell (top left) in this  $3 \times 3$  gap area is filled with the median, which is 1.58, of its surrounding  $5 \times 5$  neighbourhoods. The second cell (top middle) is filled with the median of the non-*Null* value pixels in its surrounding  $5 \times 5$  neighbourhoods including the just-filled data in the first cell, 1.58, and so on. As such, the  $3 \times 3$  gap area is filled up as shown in Figure 9.14b. If there is an even number of values within the  $5 \times 5$  neighbourhoods, the average of the middle two values is used.

#### 9.4.4 Summary of the refined pixel-to-pixel image co-registration and assessment

Incorporating the MSP, the phase correlation pixel-to-pixel image co-registration scheme is refined as below:

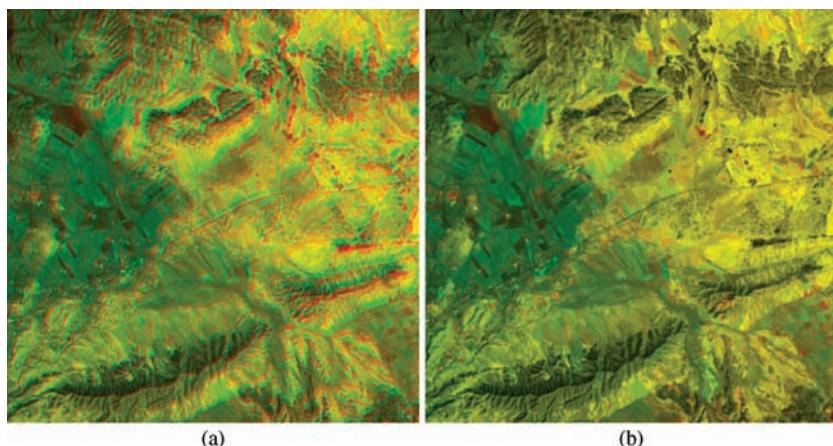
1. For the given two images for co-registration, set one as *Input* and the other *Reference* and then the

*Output* (co-registered) image frame is set based on the *Reference* image.

2. Apply SVD phase correlation to orientate the *Input* image to the *Reference* image by simple shift, rotation and minor scale change.
  3. Scan the orientated *Input* and *Reference* images using 2D fitting phase correlation with a  $64 \times 64$  calculation window to produce the images of column and line shifts between the *Input* and the *Reference* images,  $\Delta X$  and  $\Delta Y$ .
  4. Meanwhile, the RCD image recording the correlation quality between every pair of corresponding pixels of *Reference* and *Input* images is generated. The low-correlation areas in the  $\Delta X$  and  $\Delta Y$  images are masked off with a threshold of  $RCD < 0.75$ .
  5. Apply the MSP technique to fill the gaps in the masked  $\Delta X$  and  $\Delta Y$  images produced in the previous step.
  6. Use *Output-to-Input* mapping to build the *Output* image pixel by pixel based on the shift between the *Input* and the *Reference* images,  $\delta x \in \Delta X, \delta y \in \Delta Y$ .

This scheme will generate an *Output* image that is co-registered to the *Reference* image at the sub-pixel level, from the *Input* image.

Figure 9.15 presents the colour composites of a Terra-1 ASTER band 3 in green and a Landsat-7 ETM+ Pan image before and after co-registration in red. The two images covering the same area were taken from different satellite platforms with



**Figure 9.15** Red and green colour composites of (a) the original ETM+ Pan image in red and the ASTER band 3 image in green and (b) the co-registered ETM+ Pan image in red and the ASTER band 3 image in green. The crystal sharpness of (b) indicates that the ETM+ Pan image is very accurately co-registered to the ASTER band 3 image pixel by pixel and the red and green patches highlight the areas of significant spectral differences between the two images

different sensors. ETM+ is a two-way across-track scanner while ASTER is an along-track push-broom scanner. Both images are of 15 m spatial resolution but in different spectral bands: 0.52–0.9 µm for the ETM+ Pan and 0.76–0.86 µm for the ASTER band 3. There are shift, rotation and irregular geometrical variations, causing apparent mismatching between the two images, as well as spectral differences, as shown in Figure 9.15a. The ETM+ Pan image is then pixel-to-pixel co-registered to the ASTER band 3 image. The crystal sharpness of Figure 9.15b indicates a very high quality of co-registration in every part of the image where the red and green patches reveal the spectral changes or differences between the two images, illustrating the capability of the method for change detection.

Pixel-to-pixel image co-registration may fail if large parts of the images are of low correlation because of widespread spectral differences and featureless areas, or the two images are taken in very different illumination conditions. In this case, manual GCP-based co-registration is more robust as the human eye can pick up GCPs based on comprehension of textures and spatial patterns. In summary, GCP-based image warping is versatile for general purpose raster data co-registration and

georectification, whereas phase-correlation-based pixel-to-pixel image co-registration is only suitable for images with similar properties, though it achieves sub-pixel co-registration accuracy for every image pixel.

## 9.5 Summary

After a brief discussion of the major sources of geometric deformation in remotely sensed images acquired by Earth observation satellites, we introduced the details of image warping co-registration based on a polynomial deformation model derived from GCPs. The key points for this popular method are as follows:

1. The accuracy of a polynomial deformation model largely depends on the quality, number and spatial distribution of GCPs, but also on the order of the polynomials. A higher order polynomial requires more GCPs and more computing time with improved accuracy. The most commonly used polynomial deformation models are linear, quadratic and cubic.
2. The co-registered image is generated based on an output-to-input mapping procedure to

avoid pixel overlaps and holes in the output image. In this case, pixel DN resampling is necessary to draw DNs in non-integer positions in the input image for a DN in the corresponding integer position in the output image via nearest neighbour or bilinear interpolation.

Given that GCP selection is of vital importance for the accuracy and automation of image warping co-registration, we then focused on techniques for manual, semi-automatic and automatic GCP selections. Automatic GCP selection is an active research field. The key for the automation of GCP selection is the automatic detection of corner points which are good candidates for GCPs. As an example and for advanced readers, we have presented a simple and effective corner detector, based on the autocorrelation matrix.

Finally, for further reading, an optical flow image co-registration method to sub-pixel accuracy is presented in detail. The optical flow image co-registration is based on pixel-to-pixel local feature matching, which is fundamentally different from the deformation-model-based image warping methods. The phase correlation technique is used to achieve precise feature matching at sub-pixel precision. The method ensures the high quality of image co-registration at every image pixel regardless of the irregular deformations between the two images, and therefore can achieve much higher co-registration accuracy than deformation-model-based image warping co-registration methods which cannot handle irregular deformations. The pixel-to-pixel co-registration method is not, however, a georectification process and so is not versatile for dealing with raster datasets of very different properties.

## Questions

- 9.1 Describe, using a diagram, the relationship between satellite flight direction, sensor status and imaging position on the ground.
- 9.2 Give an example of how the instability of sensor status produces geometric errors.
- 9.3 What is the best sensor status that introduces minimal geometric distortion for Earth observation remote sensing?
- 9.4 Explain the relationship between focal length  $f$  and topographic distortion using a diagram.
- 9.5 Explain the relationship between imaging geometry and topographic distortion using a diagram.
- 9.6 Explain why de-skewing processing is essential for images acquired by a scanner onboard a satellite in a circular, near-polar, Sun-synchronous orbit. If a camera instead of a scanner is used for imaging, do you think that de-skewing processing is necessary and, if so, why?
- 9.7 What is output-to-input mapping (explain using a diagram) and why is it necessary for image warping co-registration?
- 9.8 How is the transformation,  $M$ , established in a polynomial deformation model?
- 9.9 How many GCPs are required to establish a linear, quadratic and cubic polynomial deformation model? Can you write down these polynomials?
- 9.10 From a diagram, derive the bilinear interpolation for pixel DN resampling. Calculate  $P_i(5.3, 4.6)$  given  $P_i(5, 4) = 65$ ,  $P_i(6, 4) = 60$ ,  $P_i(5, 5) = 72$  and  $P_i(6, 5) = 68$ .



# 10

## \*Introduction to Interferometric Synthetic Aperture Radar Techniques

In this chapter, we introduce some advanced interferometric synthetic aperture radar (InSAR) techniques for 3D terrain representation, for quantitative measurements of terrain deformation and for the detection of random land surface changes. InSAR is not normally covered by the general scope of image processing but it has become a widely used application of SAR data analysis in remote sensing. Many InSAR image processing software packages have now been developed and some popular image processing systems now include some InSAR functionality.

### 10.1 The principle of a radar interferometer

For many remote sensing applications, we use processed *multi-look* SAR images. These products represent images of the averaged intensity (or amplitude) of multiple radar looks to reduce radar speckles. The original SAR image representing all the information from the return radar signals is a single look complex (SLC) image. An SLC image is composed of complex pixel numbers which record not only the intensity (the energy of microwave signals returned from targets) but also the phase of

the signal which is determined by the distance between the target and the radar antenna.

Given a complex number of an SLC pixel,  $c = a + ib$ ,  $i = \sqrt{-1}$ , the magnitude of  $c$  is  $M_c = \sqrt{a^2 + b^2}$ , which formulates the SAR intensity image, while the phase angle of  $c$  is  $\varphi = \arctan(b/a)$ .

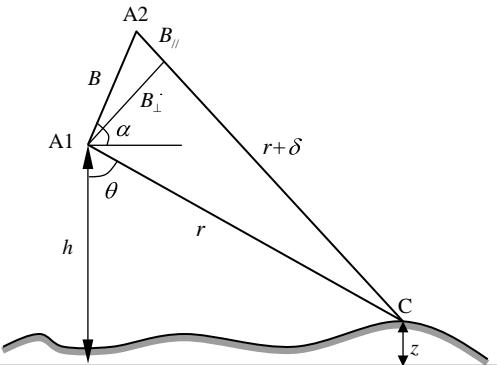
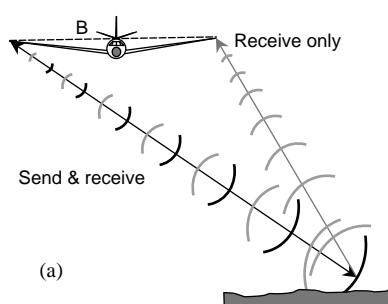
InSAR technology exploits the (commonly ignored) phase information in SAR SLC images for Earth and planetary observations. An SAR interferogram shows the phase differences between the corresponding pixels of the same object in two SAR images taken from near-repeat orbits. It represents topography as *fringes of interference*. Based on this principle, InSAR technology has been developed and used successfully for topographic mapping and measurement of terrain deformation caused by earthquakes, subsidence, volcano deflation and glacial flow (Zebker and Goldstein, 1986; Gabriel, Goldstein and Zebker, 1989; Zebker *et al.*, 1994a, 1994b; Massonnet and Adragna, 1993; Massonnet *et al.*, 1993, 1994; Massonnet, Briole and Arnaud, 1995; Goldstein *et al.*, 1993). The vertical accuracy is several tens of metres for DEM generation and at the centimetre level for measurement of terrain surface deformation.

A radar beam is nominally a single frequency electromagnetic wave. Its properties are similar to

those of monochromatic coherent light. When two nearly parallel beams of coherent light illuminate the same surface, an interferogram can be generated showing the phase shift induced by the variation of position and topography of the surface, as a result of the interference between the two beams. The same principle applies to the return radar signals. An SAR interferometer acquires two SLC images of the same scene with the antenna separated by a distance  $B$  called the *baseline*. For a single pass SAR interferometer, such as an SAR interferometer onboard an aircraft or the Space Shuttle (e.g. SRTM mission, see also Section 16.2.1), two images are acquired simultaneously via two separate antennas: one sends and receives the signals while the other receives only (Figure 10.1a). In contrast, a repeat-pass SAR interferometer acquires a single image of the same area twice from two separate orbits with minor drift which forms the baseline  $B$  (Figure 10.1b); this is the case for ERS-1 and ERS-2 SAR, ENVISAT ASAR, RADARSAT and ALOS PALSAR.

The purpose of InSAR is to derive an SAR interferogram,  $\phi$ , which is the phase difference between the two coherent SLC images (often called *fringe pair*). Firstly, the two SLC images are precisely co-registered pixel by pixel at sub-pixel accuracy based on local correlation in combination with embedded position data in SAR SLCs. The phase difference  $\phi$  between the two corresponding pixels is then calculated from the phase angles  $\varphi_1$  and  $\varphi_2$  of these two pixels through their complex numbers:

$$\phi = \varphi_1 - \varphi_2. \quad (10.1)$$



**Figure 10.2** The geometry of radar interferometry

To understand the relationship between phase difference and the InSAR imaging geometry, let us consider an SAR system observing the same ground swath from two positions, A1 and A2, as illustrated in Figure 10.2. The ground point C is then observed twice from distance  $r$  (slant range) and  $r + \delta$ . The distance difference between the return radar signals for a round-trip is  $2\delta$  and the measured phase difference  $\phi$  (interferogram) is

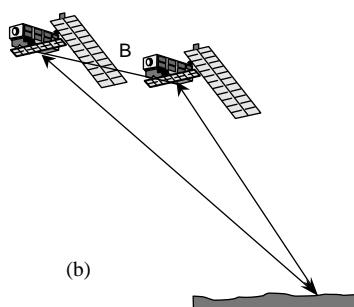
$$\phi = \frac{4\pi}{\lambda} \delta \quad (10.2)$$

or  $2\pi$  times the round-trip difference,  $2\delta$ , in radar wavelength  $\lambda$ .

From the triangle A1–A2–C in Figure 10.2, the cosine theorem provides a solution for  $\delta$  in terms of the imaging geometry as follows:

$$(r + \delta)^2 = r^2 + B^2 - 2rB \cos\left[\frac{\pi}{2} - (\theta - \alpha)\right] \quad \text{or}$$

$$(r + \delta)^2 = r^2 + B^2 - 2rB \sin(\theta - \alpha) \quad (10.3)$$



**Figure 10.1** (a) The single pass SAR interferometer with both an active antenna, sending and receiving radar signals, and a passive antenna (separated by a distance  $B$ ) for receiving signals only. (b) A repeat-pass SAR interferometer to image the same area from two visits with a minor orbital drift  $B$

where  $B$  is the baseline length,  $r$  the radar slant range to a point on the ground,  $\theta$  the SAR look angle, and  $\alpha$  the angle of the baseline with respect to the horizontal at the sensor.

The baseline  $B$  can be decomposed into two components which are perpendicular  $B_{\perp}$  and parallel  $B_{\parallel}$  to the look direction:

$$B_{\perp} = B \cos(\theta - \alpha) \quad (10.4)$$

$$B_{\parallel} = B \sin(\theta - \alpha). \quad (10.5)$$

The InSAR data processing accurately calculates the phase difference  $\phi$  between corresponding pixels between a fringe pair of SLC SAR images to produce an interferogram. The applications of InSAR are largely based on the relationships between the interferogram  $\phi$ , topography and terrain deformation, for which the baseline  $B$ , especially the perpendicular baseline  $B_{\perp}$ , plays a key role.

## 10.2 Radar interferogram and DEM

One major application of InSAR is to generate a DEM. It is clear from Figure 10.2 that the elevation of the measured point C can be defined as

$$z = h - r \cos\theta \quad (10.6)$$

where  $h$  is the height of the sensor above the reference surface (datum). This formula looks simple but the exact look angle  $\theta$  is not directly known from the SLC images. We have to find these unknowns from the data which InSAR provides. From the SAR interferogram  $\phi$ , we can express  $\delta$  by rearranging (10.2) as

$$\delta = \frac{\lambda\phi}{4\pi}. \quad (10.7)$$

Modifying (10.3) as a sine function of  $\theta - \alpha$ ,

$$\sin(\theta - \alpha) = \frac{r^2 + B^2 - (r + \delta)^2}{2rB}. \quad (10.8)$$

In this equation, the baseline  $B$  and slant range  $r$  are known and constants for both entire fringe pair

images, while the only variable  $\delta$  can be easily calculated from phase difference  $\phi$  (SAR interferogram) using Equation (10.7). Thus  $\sin(\theta - \alpha)$  is resolved.

Expressing  $\cos\theta$  in Equation (10.6) as a function of  $\alpha$  and  $\sin(\theta - \alpha)$ ,

$$\begin{aligned} z &= h - r \cos\theta \\ &= h - r \cos(\alpha + \theta - \alpha) \\ &= h - r \cos\alpha \cos(\theta - \alpha) + r \sin\alpha \sin(\theta - \alpha) \\ &= h - r \cos\alpha \sqrt{1 - \sin^2(\theta - \alpha)} + r \sin\alpha \sin(\theta - \alpha). \end{aligned} \quad (10.9)$$

In Equation (10.9), the angle of the baseline with respect to the horizontal at the sensor,  $\alpha$ , is a constant for the SAR fringe pair images and is determined by the imaging status, whereas  $\sin(\theta - \alpha)$  can be derived from the interferogram  $\phi$  using Equations (10.7) and (10.8), and the elevation  $z$  can therefore be resolved.

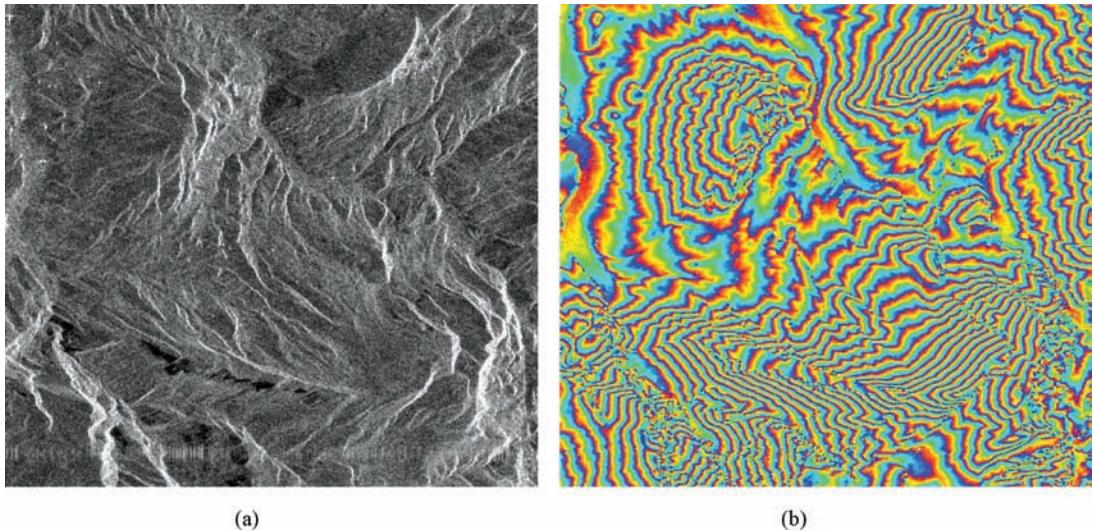
In principle, we can measure the phase difference at each point in an image and apply the above three equations based on our knowledge of imaging geometry to produce the elevation data. There is, however, a problem in this: the InSAR measured phase difference is a variable in the  $2\pi$  period, or is  $2\pi$  wrapped. Figure 10.3 shows an interferogram generated from a fringe pair of ERS-2 SAR images; the fringe patterns are like contour lines representing the mountain terrain but, numerically, these fringes occur in repeating  $2\pi$  cycles and do not give the actual phase differences which could be  $n$  times  $2\pi$  plus the InSAR measured phase difference. The phase information is recorded in the SAR data as complex numbers and only the principal phase values ( $\phi_p$ ) within  $2\pi$  can be derived. The actual phase difference should therefore be

$$\phi = \phi_p + 2n\pi. \quad (10.10)$$

Expressed in terms of the slant range difference,

$$\delta = \frac{\lambda\phi}{4\pi} = \frac{\lambda}{4\pi}(\phi_p + 2n\pi) = \frac{\lambda\phi_p}{4\pi} + \frac{n\lambda}{2}. \quad (10.11)$$

The interferometric phase therefore needs to be unwrapped to remove the modulo- $2\pi$  ambiguity



**Figure 10.3** (a) An ERS-2 SAR multi-look image; and (b) the SAR interferogram of the same scene in which the fringes, in  $2\pi$  cycles, are like contours showing the topography

so as to generate DEM data. For a perfect interferogram modulo- $2\pi$ , unwrapping can be achieved accurately via a spatial-searching-based scheme but the various decorrelation factors mean that SAR interferograms are commonly noisy. In such cases, unwrapping is an ill-portrayed problem. There are many well-established techniques for the unwrapping of noisy InSAR interferograms, each with its own merits and weaknesses, but the search for better techniques continues. The details of unwrapping are beyond the scope of this book. There are also other corrections necessary, such as the removal of the ramps caused by the Earth's curvature and by the direction angle between the two paths as they are usually not actually parallel. All the relevant functionalities are available in commercially available InSAR software packages.

We will now prove that the elevation resolution is proportional to the perpendicular baseline  $B_{\perp}$ . The partial derivative of elevation  $z$  to the slant range increment  $\delta$  is

$$\frac{\partial z}{\partial \delta} = \frac{\partial z}{\partial \theta} \times \frac{\partial \theta}{\partial \delta}. \quad (10.12)$$

From (10.2),

$$\frac{\partial z}{\partial \theta} = r \sin \theta. \quad (10.13)$$

From (10.3),

$$\frac{\partial \theta}{\partial \delta} = -\frac{r + \delta}{Br \cos(\theta - \alpha)}. \quad (10.14)$$

Assuming that  $\delta$  is very small in comparison with  $r$ ,

$$\begin{aligned} \frac{\partial z}{\partial \delta} &= r \sin \theta \left[ -\frac{r + \delta}{Br \cos(\theta - \alpha)} \right] \\ &\approx -\frac{r \sin \theta}{B \cos(\theta - \alpha)} \\ &= -\frac{r \sin \theta}{B_{\perp}}. \end{aligned} \quad (10.15)$$

We can then derive the partial derivative of elevation  $z$  to the change in phase difference (interferogram)  $\phi$ . From (10.7),

$$\frac{\partial \delta}{\partial \phi} = \frac{\lambda}{4\pi}, \quad \frac{\partial z}{\partial \phi} = \frac{\partial z}{\partial \delta} \times \frac{\partial \delta}{\partial \phi} \approx -\frac{\lambda r \sin \theta}{4\pi B_{\perp}}. \quad (10.16)$$

Therefore, for a phase increment  $\Delta\phi$ , we have

$$\Delta z = -\frac{\lambda r \sin \theta}{4\pi B_{\perp}} \Delta\phi. \quad (10.17)$$

For one fringe cycle,  $\Delta\phi = 2\pi$ , in this case,

$$\Delta z_{2\pi} = -\frac{\lambda r \sin\theta}{2B_{\perp}}. \quad (10.18)$$

The numerator in (10.18) is constant and thus the greater the value of  $B_{\perp}$ , the less the elevation increment with one cycle of  $2\pi$  and the higher the DEM resolution. Given

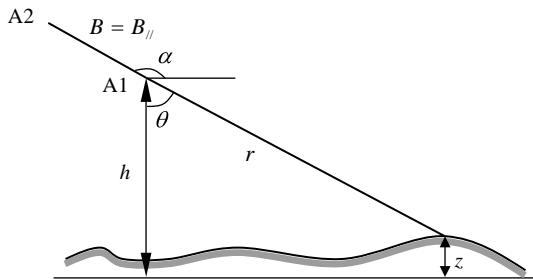
$$r = \frac{h}{\cos\theta}$$

we can rewrite (10.18) as

$$\Delta z_{2\pi} = -\frac{\lambda h \tan\theta}{2B_{\perp}}. \quad (10.19)$$

In the case of the ENVISAT SAR, where  $\lambda = 0.056$  m,  $h = 800\,000$  m, look angle  $23^\circ$  and  $B_{\perp} = 300$  m, we then have  $|\Delta z_{2\pi}| = 31.7$  m. We find that the elevation resolution of InSAR is not very high even with a 300 m perpendicular baseline. This is because the translation from phase difference to elevation is indirect, according to the geometry shown in Figure 10.2.

As hinted in Equation (10.19),  $B_{\perp}$  enables observation of the same ground position of height from different view angles. This is the key factor for DEM generation using InSAR and it is similar to DEM generation from a pair of stereo images, although it is based on quite different principles. As a simple demonstration, consider a special case,  $B_{\perp} = 0$ , where a point is observed from two SAR images at the same view angle, as shown in Figure 10.4. Then  $\delta = B_{\parallel} = B$  is a constant which is independent of the view angle  $\theta$ . In this unique case, the interferogram is a constant invariant with respect to position and it therefore contains no topographic information. An SAR fringe pair with very small  $B_{\perp}$

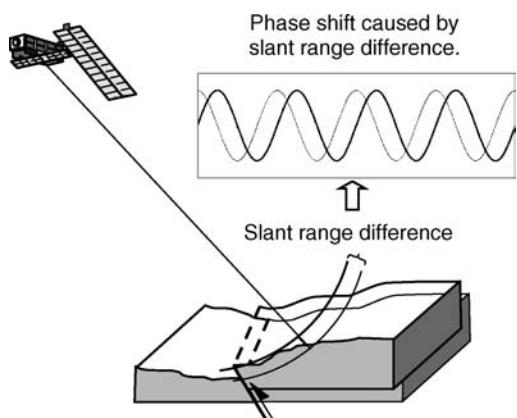


**Figure 10.4** A special case of InSAR for  $B_{\perp}=0$

is therefore insensitive to elevation. On the other hand, a very large  $B_{\perp}$ , though more sensitive to elevation, will significantly reduce the coherence level, and thus the SNR of the interferogram, because the two radar beams become less parallel. Most InSAR applications require the ratio  $B_{\perp}/r$  to be less than 1/1000. Translating this to ERS-2 and ENVISAT with an orbital altitude of about 800 km and a nominal view angle of  $23^\circ$ , the desirable  $B_{\perp}$  for effective InSAR should be less than 1000 m.

### 10.3 Differential InSAR and deformation measurement

Satellite differential InSAR (DInSAR) is an effective tool for the measurement of terrain deformation as caused by earthquakes, subsidence, glacial motion and volcanic deflation. DInSAR is based on a repeat-pass spaceborne radar interferometer configuration. As shown in Figure 10.5, if the terrain is deformed by an earthquake on a fault, then the deformation is translated directly as the phase difference between two SAR observations, made before and after the event. If the satellite orbit is precisely controlled to make the two repeat observations from exactly the same position, or at least  $B_{\perp}=0$ , as illustrated in Figure 10.4, the phase difference measured from InSAR is entirely produced by the deformation in the slant range



**Figure 10.5** Illustration of phase shift induced from terrain deformation, measured by differential InSAR

direction. This ideal zero-baseline case is, however, unlikely to be the real situation in most Earth observation satellites with SAR sensor systems. In general, the across-event SAR observations are made with a baseline  $B_{\perp} \neq 0$ , and as a result the phase difference caused by terrain deformation is jumbled with the phase difference caused by the topography. Logically, the difference between the interferogram generated from two SAR observations before terrain deformation and that from two SAR observations across the deformation event should cancel out the topography and retain the phase difference representing only terrain deformation. Topographic cancellation can be achieved from the original or from the unwrapped interferograms. The results of DInSAR can then be presented either as differential interferograms or deformation maps (unwrapped differential interferograms). For simplicity in describing the concepts of DInSAR processing, the phase difference in the following discussion refers to the true phase difference, not the  $2\pi$  wrapped principal phase value.

With two pre-event SAR images and one post-event image (within the feasible baseline range), a pre-event and a cross-event fringe pair can be formulated, so a differential interferogram can be directly derived. The DInSAR formula, after correction for flattening of the Earth's curvature, is as follows (Zebker *et al.*, 1994a):

$$\delta\phi = \phi_{2\text{flat}} - \frac{B_{2\perp}}{B_{1\perp}} \phi_{1\text{flat}} = \frac{4\pi}{\lambda} D \quad (10.20)$$

$$\phi_{\text{flat}} = \frac{4\pi}{\lambda} (\theta - \theta_0) \cdot B_{\perp} \quad (10.21)$$

where  $\theta_0$  is the look angle to each point in the image assuming zero local elevation on the ground (the reference ellipsoid of the Earth);  $\phi_{1\text{flat}}$  and  $\phi_{2\text{flat}}$  are the unwrapped pre-event and cross-event interferograms after the flattening correction for the Earth's curvature; and  $D$  represents the possible displacement of the land surface.

The ratio between the perpendicular baselines of the two fringe pairs is necessary because the same  $2\pi$  phase difference represents different elevations depending on  $B_{\perp}$  according to formula (10.19).

The operation cancels out the stable topography and reveals the geometric deformation of the land

surface. Formula (10.20) indicates that the deformation  $D$  is directly proportional to the differential phase difference  $\delta\phi$ , thus DInSAR can provide measurements of terrain deformation at better than half the wavelength of SAR at millimetre accuracy. For instance, the wavelength of the C-band SAR onboard ENVISAT is 56 mm and thus 28 mm deformation along the slant range will be translated to  $2\pi$  phase difference in a cross-event C-band SAR interferogram.

As an alternative approach, if a high-quality DEM is available for an area under investigation,  $\phi_1$  can be generated from the DEM with an artificially given baseline equal to the baseline of the across-event fringe pair and simulated radar imaging geometry (Massonnet *et al.*, 1993, 1994; Massonnet, Briole and Arnaud, 1995). In this case, DInSAR is a simple difference between the cross-event SAR interferogram  $\phi_2$  and the simulated interferogram of topography  $\phi_{1\text{sim}}$ :

$$\delta\phi = \phi_2 - \phi_{1\text{sim}} = \frac{4\pi}{\lambda} D \quad (10.22)$$

The advantages of using a DEM to generate  $\phi_{1\text{sim}}$  are that it is not restricted by the availability of suitable SAR fringe pairs and that the quality of  $\phi_{1\text{sim}}$  is guaranteed without the need for unwrapping. As further discussed in the next section, the quality of an SAR interferogram is often significantly degraded by decoherence factors and, as a result, unwrapping will unavoidably introduce errors.

Obviously, one crucial condition for DInSAR is that the satellite position for SAR imaging is controlled to a high precision and is frequently recorded to ensure accurate baseline calculation. If the satellite can be controlled to repeat exactly, providing an identical orbit, then the so-called zero-baseline InSAR is achieved, which, without topographic information, is in fact the same as a DInSAR measurement of deformation directly. In many applications, it is not always necessary to go through this rather complicated process to generate differential SAR interferograms. Since the fringe patterns induced from topography and from terrain deformation are based on different principles, they are spatially very different and can often be visually separated for qualitative analysis. Also, since the terrain-deformation-induced fringes

are a direct translation from deformation to phase difference, they are often localized and show significantly higher density than the topographic fringes, especially in an interferogram with a short  $B_{\perp}$ . For a flat area without noticeable topographic relief, any obvious fringe patterns in a cross-event SAR interferogram should be the result of terrain deformation. In such cases, we can use InSAR for the purpose of DInSAR with great simplicity.

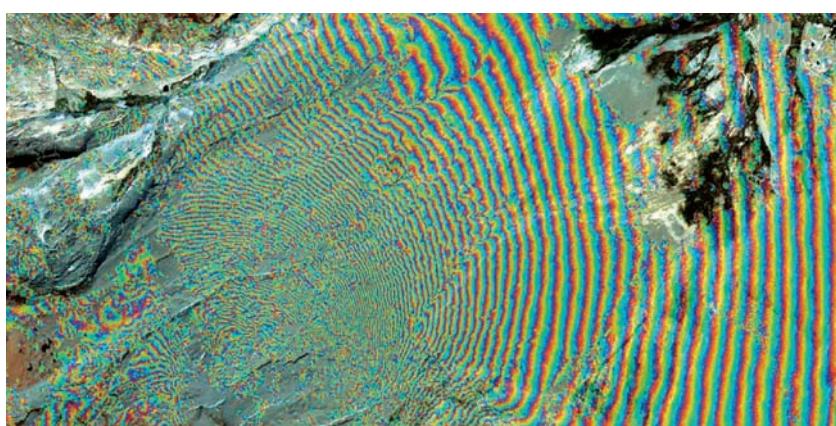
Figure 10.6 is an ENVISAT ASAR interferogram (draped over a Landsat-7 ETM+ true colour composite) showing the terrain deformation caused by an earthquake of moment magnitude 7.3 which occurred in the Siberian Altai, on 27 September 2003. The high-quality fringes are produced mainly in the basin where elevation varies gently over a range of less than 250 m. With  $B_{\perp} = 168$  m for this fringe pair, of wavelength  $\lambda = 56$  mm, orbital altitude  $h = 800$  km and look angle  $\theta = 23^\circ$ , we can calculate from Equation (10.19) that each  $2\pi$  phase shift represents 56.6 m elevation. Thus the 250 m elevation range translates to no more than five  $2\pi$  fringes. The dense fringe patterns in this interferogram are dominantly produced by the earthquake deformation.

Certainly, when we say that DInSAR can measure deformation on a millimetre scale, we mean the average deformation in the SAR slant range direction in the image pixel, which is about 25 m spatial resolution for ERS InSAR.

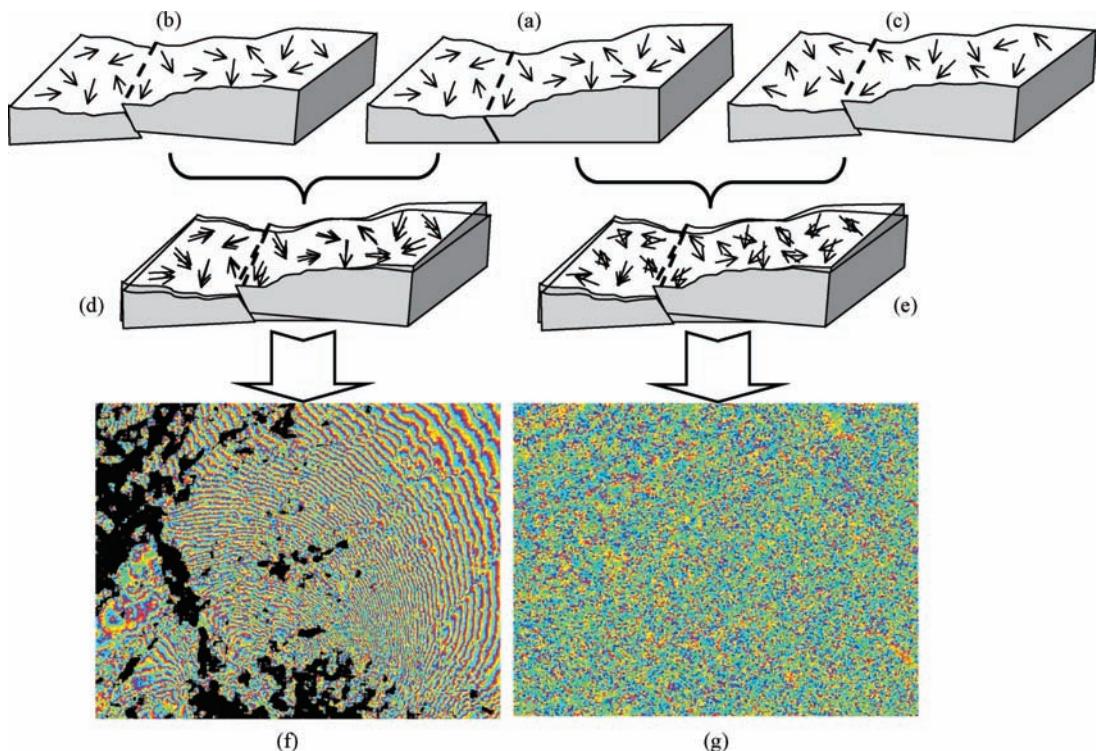
## 10.4 Multi-temporal coherence image and random change detection

So far the application of repeat-pass InSAR is considerably restricted from any areas subject to continuous random changes such as in dense vegetation cover. To generate a meaningful interferogram, two SAR images must be highly coherent or have high coherence. The InSAR phase coherence is a measurement of local spatial correlation between two SAR images. Any random changes to the scatterers on the land surface between the two acquisitions will result in irregular variations in phase which will reduce the coherence. In particular, the random variation exceeding half a wavelength of the radar beam in the slant range direction will cause a total loss of coherence.

As shown in Figure 10.7, the terrain deformation caused by earthquakes and so on typically comprises a block 3D shift which largely does not alter the surface scatterers randomly and, therefore, as long as there are no other factors causing random changes on the land surface, the phases of return SAR signals from ground scatterers will all shift in the same way as the block movement; this collectively coherent phase shifting can be recorded as a high-quality SAR interferogram giving quantitative measurements of the deformation (Figure 10.7d). If, however, the land surface is subject to random changes, the



**Figure 10.6** An ENVISAT interferogram draped over an ETM true colour composite showing the terrain deformation produced by an earthquake of moment magnitude 7.3 which occurred in the Siberian Altai on 27 September 2003



**Figure 10.7** Illustration of the effects of terrain deformation and random surface changes on interferograms. The small arrows in diagrams (a) to (c) represent the phase angles of return SAR signals from scatterers on the land surface. The block movement along a fault between SAR acquisitions (a) and (b) results in a coherent phase shift in the same direction as the motion between the two blocks (d), but there are no random phase shifts within each of the faulting blocks and thus a good interferogram (f) is produced which records the terrain deformation. In the case where random changes are involved in addition to the faulting motion, as shown between (a) and (c), random phase shifts which are not related to the fault movement are introduced and the return signal phases between the two acquisitions are no longer coherent (e). As a result, the interferogram (g) is chaotic and does not show any meaningful interferometric patterns

phases of scatterers are altered randomly and will no longer be collectively coherent; the coherence is consequently lost and the interferogram becomes chaotic (Figure 10.7e).

Any land surface is subject to continuous random change caused by many decorrelation factors such as vegetation growth, wind-blown sands and erosion. Since random changes on a land surface are cumulative, an SAR fringe pair with a long temporal separation is likely to have low coherence and cannot therefore be used to produce a high-quality interferogram. The low coherence means a lack of interferometric information but this may still be useful in indicating random changes on the land surface, although it cannot give direct

information of the nature of those changes. With the development of the InSAR technique and applications, the value of multi-temporal coherence imagery as an information source for surface change detection has been widely recognized and many successful application cases have been reported (Corr and Whitehouse, 1996; Ichoku *et al.*, 1998; Liu *et al.*, 1997a; Liu, Lee and Pearson, 1999; Lee and Liu, 1999; Liu *et al.*, 2001; Liu *et al.*, 2004).

Phase coherence of two SLC SAR images represents the local correlation of the radar reflection characteristics of the surface target between two observations. This can be estimated by ensemble averaging  $N$  neighbouring pixels of complex SAR

data as

$$\rho = \frac{\left| \sum_{l=1}^N z_1 l z_2^* \right|}{\left( \sum_{l=1}^N z_1 l z_1^* \sum_{l=1}^N z_2 l z_2^* \right)^{1/2}} \quad (10.23)$$

where  $\rho$  is the coherence magnitude estimate,  $z_1, z_2$  the complex values of the two SAR images, and  $*$  the complex conjugate.

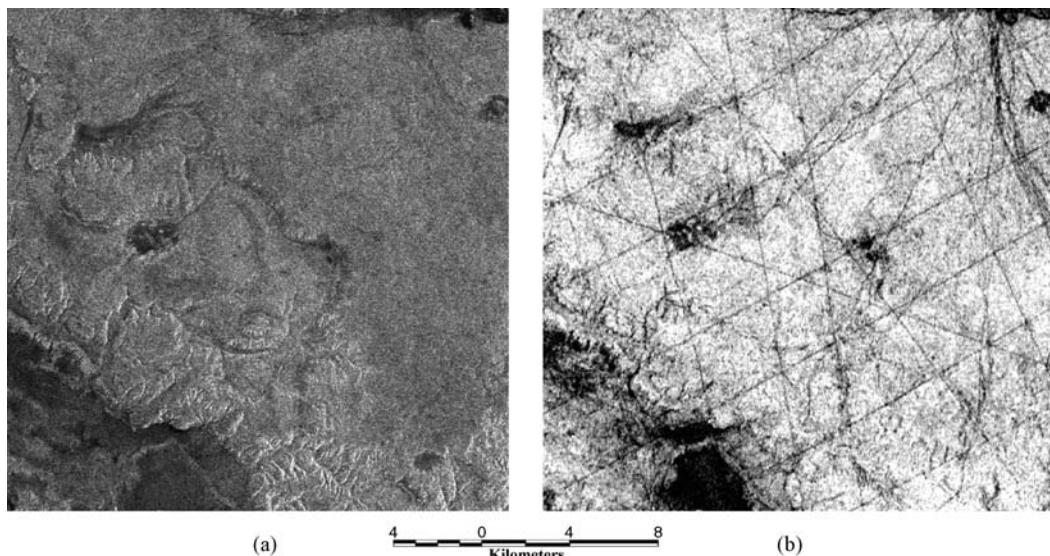
Obviously, the value range of  $\rho$  is  $[0, 1]$ , going from totally incoherent to completely coherent.

Besides providing a quality assessment for SAR interferograms, the interferometric coherence technique has been widely used for the detection of surface random change phenomena such as rapid erosion, human-activity-induced land disturbance and assessment of earthquake destruction. Figure 10.8 presents an application example showing an ERS-2 SAR multi-look image (a) of an area in the Sahara Desert and a coherence image (b) of the area derived from two SAR acquisitions with a temporal separation of 350 days. The coherence image reveals a mesh of dark straight lines which represent seismic survey lines made during the 350 days between the two SAR acquisitions,

while the multi-look image of the second acquisition of this InSAR pair shows nothing. For the seismic survey, trenches were dug to place the seismic sensors and, since then, the land surface has recovered. Since the land surface appears as before, we can see no sign of these seismic survey lines in the multi-look image. The land surface materials along the survey lines are, however, significantly disturbed and as a result their scattering properties are altered randomly causing the loss of coherence.

## 10.5 Spatial decorrelation and ratio coherence technique

Apart from temporal random land surface changes during the period between two SAR acquisitions, there are other factors which can cause the loss of coherence in InSAR coherence images. Among these decorrelation factors, baseline and topographic factors are often called spatial decorrelation as they are relevant to the geometric relations of sensor position and target distribution. In terrain with a considerable relief, topographic decorrelation due to slope angle is often the dominant decorrelation



**Figure 10.8** An ERS SAR multi-look image (a) and an InSAR coherence image generated from two SAR acquisitions with 350 days of temporal separation (b) of an area in the Sahara Desert. The coherence image reveals a mesh of seismic survey lines made during the period between the two acquisitions. These seismic lines are not shown in the multi-look image formed from the second acquisition of this InSAR pair

factor. This type of decorrelation is an intrinsic property of a side-looking and ranging SAR system. The decorrelation is overwhelming, particularly on a foreshortened or layover slope where the coherence drops dramatically towards zero. The low-coherence features of such slopes can easily be misinterpreted in the coherence imagery as an unstable land surface subject to rapid random change, even on a highly stable slope. It is important to separate topographic spatial decorrelation from the temporal decorrelation to achieve effective detection of random land surface changes.

Phase coherence decreases with the increase of  $B_{\perp}$  as characterized in the formula (Zebker and Villasenor, 1992)

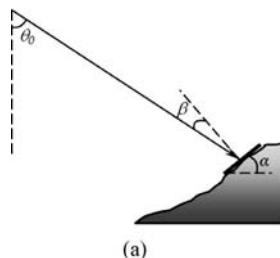
$$\rho_{\text{spatial}} = 1 - \frac{2 \cos \beta R_y}{\lambda r} B_{\perp} \quad (10.24)$$

where  $\beta$  is the incidence angle,  $R_y$  the ground range resolution,  $\lambda$  the radar wavelength and  $r$  the distance from the radar sensor to the centre of a resolution element (slant range).

In general, SAR fringe pairs with small  $B_{\perp}$  are desirable for coherence image applications. From formula (10.24) we can further prove that the influence of spatial decorrelation varies with topography.

Let  $\theta_0$  represent the nominal look angle of the SAR ( $23^{\circ}$  for ERS-2) and  $\alpha$  the local terrain slope measured upwards from the horizon away from the radar direction; then  $\beta = \theta_0 - \alpha$  (Figure 10.9a). The ground range resolution is thus a function of the local terrain slope:

$$R_y = \frac{c}{2B_w |\sin(\theta_0 - \alpha)|} \quad (10.25)$$



(a)

where  $c$  is the speed of light and  $B_w$  the frequency bandwidth of the transmitted chirp signal.

The ground range resolution  $R_y$  increases rapidly when the surface is nearly orthogonal to the radar beam and becomes infinite if the terrain slope is equal to the nominal look angle (i.e.  $\alpha = \theta_0$ ) (Figure 10.9b). Note that  $R_y$  is practically limited to a finite value because the terrain is not an infinite plane. The effect of a large value of  $R_y$  on the decorrelation is, however, significant in the case of the surface slope facing the radar. Substituting (10.25) into (10.24) results in a modified spatial decorrelation expression, as a function of perpendicular baseline and terrain slope (topography) (Lee and Liu, 2001):

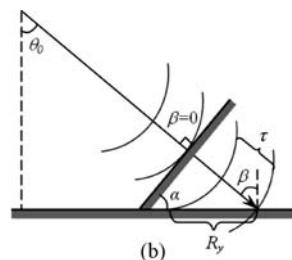
$$\rho_{\text{spatial}} = 1 - AB_{\perp} |\cot(\theta_0 - \alpha)| \quad (10.26)$$

where  $A = c/\lambda r B_w$ , a constant for the SAR system.

This spatial decorrelation function describes the behaviour of topographic decorrelation as well as baseline decorrelation. For a given baseline, the correlation decreases as the local terrain slope approaches the value of the nominal incidence angle, while the increase in baseline will speed up deterioration of the correlation.

We now introduce a method of ratio coherence for analysing and separating spatial and temporal decorrelation phenomena. The approximate total observed correlation of the returned radar signals can be generalized as a product of temporal and spatial correlation:

$$\begin{aligned} \rho &= \rho_{\text{temporal}} \cdot \rho_{\text{spatial}} \\ &= \rho_{\text{temporal}} \cdot (1 - AB_{\perp} |\cot(\theta_0 - \alpha)|). \end{aligned} \quad (10.27)$$



(b)

**Figure 10.9** (a) The relationship between nominal radar look angle  $\theta_0$ , radar incident angle  $\beta$  and terrain slope angle  $\alpha$ ; and (b) given the nominal radar look angle  $\theta_0$ , the incident angle  $\beta$  and the ground range resolution  $R_y$  are decided by the terrain slope angle  $\alpha$ . When the slope is perpendicular to the incident radar beam or  $\beta = 0$ , the range resolution  $R_y$  is indefinite

Consider three SAR observations named 1, 2 and 3 in time sequence. A ratio coherence image can then be established by dividing a coherence image with long temporal separation by another with relatively short temporal separation as

$$\begin{aligned}\frac{\rho_{\text{spatial}}^{13}}{\rho_{\text{spatial}}^{12}} &= \frac{\rho_{\text{temporal}}^{13}}{\rho_{\text{temporal}}^{12}} \cdot \frac{\rho_{\text{spatial}}^{13}}{\rho_{\text{spatial}}^{12}} \\ &= \frac{\rho_{\text{temporal}}^{13}}{\rho_{\text{temporal}}^{12}} \cdot \frac{1 - AB_{\perp}^{13}|\cot(\theta_0 - \alpha)|}{1 - AB_{\perp}^{12}|\cot(\theta_0 - \alpha)|}\end{aligned}\quad (10.28)$$

where the superscripts represent the image pair for each coherence image.

Alternatively, we can state that the total ratio of coherence consists of a temporal ratio part and a spatial ratio part as

$$\eta = \eta_{\text{temporal}} \cdot \eta_{\text{spatial}}. \quad (10.29)$$

From the above formulae, we have the following observations.

1. Obviously, the temporal ratio part always satisfies  $\eta_{\text{temporal}} \leq 1$  because the temporal change is a cumulative process and  $\rho_{\text{temporal}}^{13} \leq \rho_{\text{temporal}}^{12}$  is always true as the temporal separations  $\Delta T^{13} > \Delta T^{12}$ .
2. If the baselines are  $B_{\perp}^{13} \geq B_{\perp}^{12}$ , then  $\rho_{\text{spatial}}^{13} \leq \rho_{\text{spatial}}^{12}$  and thus the spatial ratio  $\eta_{\text{spatial}} \leq 1$  and the total ratio  $\eta \leq 1$  for all slopes.

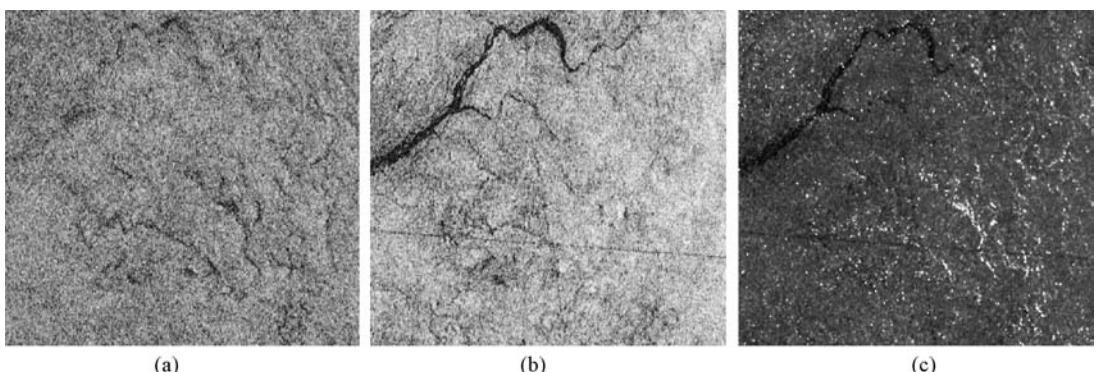
3. For the case of  $B_{\perp}^{13} < B_{\perp}^{12}$ , then  $\rho_{\text{spatial}}^{13} > \rho_{\text{spatial}}^{12}$  and the spatial ratio  $\eta_{\text{spatial}} > 1$  in general. The difference between  $\rho_{\text{spatial}}^{13}$  and  $\rho_{\text{spatial}}^{12}$  will become significant when the terrain slope is nearly normal to the incident radar beam (i.e. the slope angle approaches the radar look angle) and the spatial ratio will be abnormally high ( $\eta_{\text{spatial}} \gg 1$ ) producing high total ratio coherence  $\eta \gg 1$ . However, for the areas other than direct radar-facing slopes, we would have  $\rho_{\text{spatial}}^{13} \approx \rho_{\text{spatial}}^{12}$  and  $\eta_{\text{spatial}} \approx 1$ .

Item 3 above specifies the feasible working condition for the ratio coherence:

$$\eta = \frac{\text{Coherence of large } \Delta T \text{ and short } B_{\perp}}{\text{Coherence of small } \Delta T \text{ and long } B_{\perp}}. \quad (10.30)$$

The formula (10.30) specifies a ratio coherence image in which the numerator coherence image has a longer temporal separation and shorter baseline than the denominator coherence image. This ratio coherence image provides effective identification and separation of decoherence features sourced from spatial and temporal decorrelation, as itemized below (Figure 10.10):

1. Areas of total topographic decorrelation along the radar-facing slopes are highlighted as



**Figure 10.10** Generation of a ratio coherence image. (a) *Coh1*: Coherence image of 35 days and  $B_{\perp} = 263$  m (short temporal separation with long baseline). (b) *Coh2*: Coherence image of 350 days and  $B_{\perp} = 106$  m (long temporal separation with short baseline). (c) The ratio coherence image *Coh2/Coh1*: bright features represent topographic decorrelation on steep slopes facing the radar illumination, while dark features are detected random changes over a largely stable background in grey

- abnormally bright features because  $\eta_{\text{spatial}} \gg 1$  and then  $\eta \gg 1$ .
2. The temporal decorrelation of random changes on the land surface appears as dark features because  $\eta_{\text{temporal}} < 1$  and  $\eta_{\text{spatial}} \approx 1$  for areas not subject to severe topographic decorrelation and thus  $\eta < 1$ .
  3. The stable areas, not subject to either temporal or spatial decorrelation, appear as a grey background where  $\eta \approx 1$  as  $\eta_{\text{temporal}} \approx 1$  and  $\eta_{\text{spatial}} \approx 1$ .

## 10.6 Fringe smoothing filter

A repeat-pass SAR interferogram is often noisy because of reduced coherence caused by temporal and spatial decorrelation, as shown in Figure 10.11a. The quality of an interferogram can be considerably improved if we can reduce noise by smoothing. Since the interferogram  $\phi$  is  $2\pi$  wrapped (Figure 10.12a), ordinary smoothing filters are not directly applicable to the reduction of noise in such discontinuous periodic data. Here we introduce a simple phase fringe filtering technique (Wang, Prinet and Ma, 2001) which has the following steps:

1. For the interferogram  $\phi$ ,  $\sin\phi$  and  $\cos\phi$  are continuous functions of  $\phi$ , as shown in Figure 10.12b. Thus we convert the original interferogram into images of  $\sin\phi$  and  $\cos\phi$ . All the trigonometric functions are wrapped in a

cycle of  $\pi$ . Within the  $2\pi$  cycle of  $\phi$ , we can only retrieve angles in the range of  $[0, \pi/2]$  and  $[\pi, 3\pi/2]$  from  $\sin\phi$ , and  $[0, \pi]$  from  $\cos\phi$ . The combination of  $\sin\phi$  and  $\cos\phi$ , through a tangent function, allows us to retrieve the phase angle  $\phi$  in its original  $2\pi$  cycle.

2. A smoothing filter can then be applied to these two images of  $\sin\phi$  and  $\cos\phi$ .
3. Retrieval of the filtered interferogram  $\bar{\phi}$  from the smoothed filtered  $\overline{\sin\phi}$  and  $\overline{\cos\phi}$  is performed by

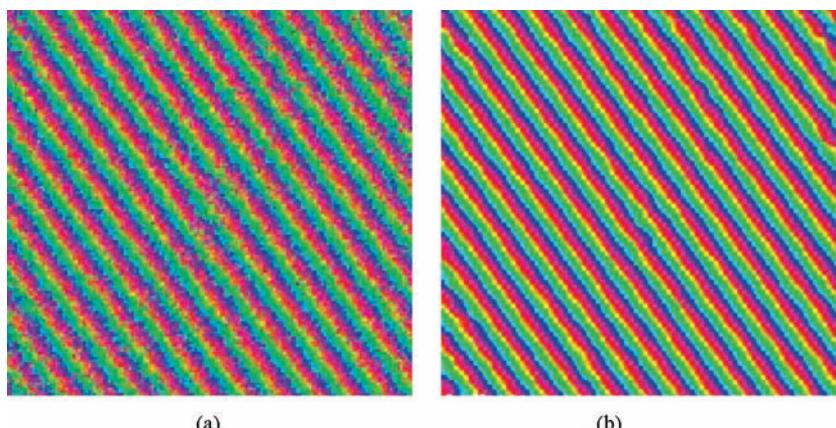
$$\bar{\phi} = \arctan\left(\frac{\overline{\sin\phi}}{\overline{\cos\phi}}\right).$$

Here the signs of  $\overline{\sin\phi}$  and  $\overline{\cos\phi}$  dictate the quadrant of the smoothed phase angle  $\bar{\phi}$ . The smoothed phase angle  $\bar{\phi}$  is within  $0-\pi/2$  for  $(+, +)$ ;  $\pi/2-\pi$  for  $(-, +)$ ;  $\pi-3\pi/2$  for  $(-, -)$ ; and  $3\pi/2-2\pi$  for  $(+, -)$ .

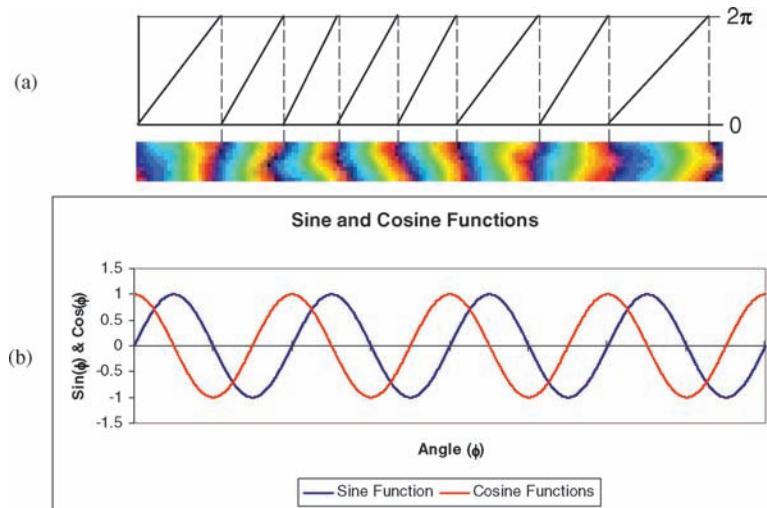
The window size of the smoothing filter used must be small compared with the half wavelength of  $\sin\phi$  and  $\cos\phi$  in the image of the interferogram. Figure 10.11b is a filtered interferogram with a  $5 \times 5$  mean filter showing significantly reduced noise.

## 10.7 Summary

Several InSAR techniques are introduced in this chapter. It is very important to notice that different



**Figure 10.11** Illustration of fringe smoothing filtering: (a) the original interferogram is quite noisy because of temporal and spatial decorrelation; and (b) after fringe smoothing filtering, the interferogram is cleaner and clearer than the original



**Figure 10.12** (a) An interferogram  $\phi$  is a discontinuous periodic function in the  $2\pi$  cycle; and (b) functions of  $\sin \phi$  and  $\cos \phi$  are also  $2\pi$  periodic but are continuous

techniques apply to different applications, as outlined below:

- *InSAR interferogram is used for DEM generation:* The single pass configuration of InSAR is preferred for the complete elimination of temporal decorrelation and to ensure a high-quality interferogram. The wider the perpendicular baseline, the higher the elevation resolution. A wider  $B_{\perp}$  introduces more severe spatial decorrelation which degrades the quality of the interferogram and then the DEM. Usually  $B_{\perp}$  of a few tens to a few hundreds of metres is used. The relationship between InSAR and topography is established by the geometry of the slightly differing view angles between the two observations. The elevation resolution of InSAR is, therefore, for C-band SAR, for example, no better than 10 m in general.
- *Differential InSAR (DInSAR) is used for the measurement of terrain deformation:* This uses repeat-pass InSAR with at least one pair of cross-deformation event SAR acquisitions. The differential phase difference is directly translated from the deformation magnitude measured at  $2\pi$  for half radar wavelength (28 mm for C-band SAR). The technique is therefore very sensitive to terrain deformation and can achieve millimetre-scale accuracies. This millimetre-scale accuracy is, however, an average deformation over an image pixel which is, for ERS InSAR, about 25 m. A short  $B_{\perp}$  is preferred for DInSAR to minimize fringes of topography. Ideally, a zero-baseline InSAR configuration would replace DInSAR.
- *InSAR coherence technique is used for random change detection:* It is very important to notice that this technique is for detection rather than measurement. A coherence technique must be based on repeat-pass InSAR. The random land surface changes are cumulative with time and this reduces coherence in a coherence image derived from two SAR SLC images with temporal separation and which can then be detected as low-coherence features. As coherence is affected by both temporal and spatial (topographic) decorrelation, a short  $B_{\perp}$  is preferred to minimize the spatial decorrelation, thus to achieve effective temporal change detection.
- *Ratio coherence technique is developed to separate spatial and temporal decorrelation:* This technique is defined as the ratio of a coherence image with long temporal separation and short  $B_{\perp}$  divided by a coherence image with short temporal separation and long  $B_{\perp}$ . Such a ratio coherence image presents spatial decorrelation in bright tones and temporal decorrelation of random changes in dark tones on a grey background representing the stable land surface.

## Questions

- 10.1 Describe, using diagrams, the basic configuration of single path and repeat-path InSAR.
- 10.2 Use a diagram to explain the principle of InSAR.
- 10.3 What are the differences between InSAR, DInSAR and coherence techniques, and what are their main applications?
- 10.4 In comparison with photogrammetric methods using stereo imagery, discuss the advantages and disadvantages of using InSAR for DEM generation.
- 10.5 Select the only correct statement from the four below and explain the reasons for your choice:
  - (a) The InSAR technique can be used to produce DEMs of centimetre-level accuracy of elevation.
  - (b) The differential InSAR technique can be used to produce DEMs of centimetre-level accuracy of elevation.
  - (c) The differential InSAR technique can detect up to centimetre-level (half radar wavelength) random changes of land surface.
  - (d) An SAR coherence image can detect up to centimetre-level (half radar wavelength) random changes of land surface.
- 10.6 Describe the ratio coherence technique and explain the conditions of baseline and temporal separation necessary for this technique to be effective.
- 10.7 Why can ordinary smoothing filters not be applied directly to an interferogram image? Describe the procedure of fringe smoothing filtering.

# **Part Two**

## **Geographical Information Systems**



# 11

## Geographical Information Systems

### 11.1 Introduction

The geographical information system or GIS needs very little introduction these days since it is a massively expanding and developing technology, and it is certainly a mature one today. There are many aspects of GIS which we now take for granted, such as online route finding tools which use postcode gazetteers to identify geographic features and addresses, or in-car navigation systems. The appearance of the likes of Google Earth has transformed people's perception of their environment and, possibly, of what we geospatial scientists do for a living. Essentially, people are thinking far more geographically than ever before. GIS is very soon to become part of the GCSE school curriculum in the United Kingdom, and this will help to increase awareness at a much earlier age. It should not be too long before we see a marked increase in the number of people embarking on geospatial careers, or perhaps just taking geospatial information even more for granted.

Software tools become more sophisticated and easier to use, as well as more effective and faster. Many processes that can be done now at the simple click of a button, only a few years ago would have taken a little while to set up and execute. This is of course advantageous but there is a temptation to trivialize what is being achieved. Underneath all the many impressive things that software tools can now do, on your desk, in your car, on the Internet, in

providing you with vast quantities of information, are still the same simple operations and processes and they are no less important. The data they access are, however, more voluminous, more efficiently organized, more effectively processed and more widely available. There is perhaps also a danger in not being aware of what happens behind that simple button click, of misunderstanding or misusing the technology, or at least of not being aware of a mistake. In this respect, the roles of human critique and of simple visualization become more important, not less, since a critical perspective is essential to ensure the quality or at least the reliability and relevance of the result.

One of the principal tools in the hands of remote sensing scientists is image processing. It is not the only tool but is one which fundamental at some stage of a remote sensing application project of some kind, especially since it is a vital contributor to the visualization of remotely sensed imagery. Visualization is also a vital part of any digital analysis in GIS; written reports are important but rather impotent without the ability actually to see the results. A great many image processing procedures lie at the heart of GIS, mainly in raster GIS of course, but there are also parallels in vector GIS. We have tried to point out the links between the various aspects of GIS described here and their relevant image processing counterparts (described in Part One) wherever possible, as well as linking to case studies (in Part Three) in which the techniques have been applied.

There are many excellent and comprehensive published texts on GIS, as listed in the general references, so this part of our book is not intended as a general GIS textbook but aims to summarize the techniques and principles which are useful in remote sensing applications within or connected with ‘geoscientific’ fields. We refer here to the many preparatory tasks as well as to the spatial analysis of data within this geoscientific context. Of course, that does not imply that the tasks dealt with here are the exclusive domain of the geosciences, but we stress that some of the more advanced tasks and techniques are perhaps (arguably) not generally within the scope of GIS use by the general geographical public.

This seems a convenient point to clarify our definition of the acronym ‘GIS’ in this book. In some instances it is taken to stand for *Geographic Information Systems* or *Geographical Information Science* and, in other instances, *Geographic Information Systems and Science*; none of these is incorrect as such, but without feeling it necessary to make the whole thing sound more impressive than it already is, and to make things simple, we will stick with *Geographical Information System*. The descriptions of processes and procedures in this part of the book are largely generic and do not therefore represent the practice of any particular software suite or the theory of any particular individual, but where certain aspects are relevant to specific GIS software packages or published research, these are stated.

## 11.2 Software tools

In the light of the conceptual overlap that is the subject of this book, it should also be understood that there are a great many other software tools, used in related disciplines, which have elements in common but do not generally qualify as ‘GIS’. Such tools involve spatially referenced information and perform very similar operations on raster and vector objects but we tend not to refer to them in the same breath, perhaps because generally speaking they are in use within far more specific application areas; one of the great strengths of GIS is its status as a general purpose tool.

Such software suites include ER Mapper (a sophisticated and totally transparent image processing

engine); Geosoft (a raster processing suite containing tools for the processing of geophysical data, and the production of maps, used largely by the mining/exploration industry); Micromine (a vector handling package for managing and displaying sub-surface and surface point geochemical and borehole data, again used primarily by the mining/exploration industry); Petrel and Landmark (suites used almost entirely by the petroleum industry, for the processing, analysis and visualization of surface and sub-surface raster and vector data); and Surfer (a sophisticated and generic toolset for gridding of vector data to produce raster grids). There are many more.

One of the limitations of GIS in geosciences lies in the fundamental concept of the *layer*. The Earth’s surface is a conceptual boundary for GIS: i.e., we cannot adequately describe sub-surface objects or the relationships between them. Sub-surface geological phenomena are, for most purposes, excluded from GIS except in that they exist as discrete features. For example, sub-surface horizon maps can be treated like any other spatial data layer, but features which intersect one layer cannot be made to intersect another in a geologically meaningful way. This is partly because the separation between layers is an arbitrary distance and is not deemed to be significant for the purposes of most GIS operations and procedures. Fortunately, however, there are other software suites which do allow for such concepts and provide more complete 3D functionality for the geosciences.

We have provided a list of GIS software packages and the URLs of their websites in Appendix B; the list includes both proprietary and shareware tools.

## 11.3 GIS, cartography and thematic mapping

Is there a difference between a GIS and a conventional cartographic map? A map is of course a GIS in itself, i.e. it is an analogue spatial database which requires perception and interpretation to extract the embedded information. Once on paper, however, it is fixed and cannot be modified. A GIS display of a map, on the other hand, is inherently flexible. Unlike the conventional paper map, it does not require every piece of information to be visible

at the same time. It can also change the depiction of a particular object according to the value of one of its attributes. Let's not forget, of course, that the cartographic map is still also a vital analogue product of our digital analysis!

This ability to display information selectively according to a particular objective is known as *thematic mapping*. Thematic maps are commonplace in atlases, textbooks and other published journal articles; standard topographic maps for instance are thematic. Thematic maps can be divided into two groups: qualitative and quantitative. Qualitative maps show the location or distribution of nominal data and in many respects qualitative thematic maps are similar to general purpose maps, except that they focus on specific themes. Quantitative maps show the variations or changing magnitudes of spatial phenomenon in different places; and quantitative maps illustrate data that can be processed into ordinal, interval or ratio scales. The thematic map is therefore a very basic component and product of remote sensing applications that involve image processing and/or GIS. The case studies described in Part Three all involve production of some kind of thematic product, either as the end result or as an intermediary stage of a broader analysis.

## 11.4 Standards, interoperability and metadata

With data and software growing in volume and capability every day, there follows the increasing need to be transparent and rigorous in the recording of quality, provenance and validity, as matters of best practice in GIS. The Open Geospatial Consortium (OGC) has been formed in relatively recent years as an internationally coordinated volunteer organization (in which there are several hundred member companies, agencies and universities) which is responsible for the driving of standards in interoperability and quality within the geospatial community (in its broadest sense). A wealth of information can be found on the OGC website, describing OpenGIS technical standards and specifications, and best practices for data, metadata and procedures ([www.opengeospatial.org/standards/](http://www.opengeospatial.org/standards/)) within all aspects of GIS.

Many proprietary software suites now incorporate open application programming interfaces to allow users to customize and develop tools for their working environment, both locally and for communication via the wired and wireless Internet. The current trend in software development in a growing international market is towards scalable and modular products, so allowing users to customize the tools according to their own needs. A parallel trend is in the sharing of technological development, with highly specialized third-party modules from one product being incorporated (as 'plug-ins', for instance) into the main suite of another; GIS has now entered the world of 'plug and play'!

The improvement of interoperability and standards is one of the great 'reliefs' as GIS comes of age. Moving information, integrating it, sharing it and recording its provenance and quality demand openness about its format and structure, how it was created and what has been done to it subsequently. Being able to import data is vital, but so too is the ability to export and share data. Thankfully, there is far less difficulty in satisfying these needs now and the trend in interoperability is ongoing and entirely positive.

Metadata describe many aspects of geospatial information, especially the content, quality and provenance of the data. This information is vital to the understanding of the information by others who may use the data. The main uses are for general database organization and for digital catalogues and Internet servers. Such information is usually stored in a standard form, such as *.xml*, and so can be created or edited using a standard text editor program. The recording of metadata seems such an easy and trivial thing that it is often overlooked or neglected, but with the growing volumes of data at our fingertips it becomes more and more of a necessity. There are several basic categories of metadata which should be recorded, as follows:

- General identification – the name and creator of the data, the general location of the data, date of creation and any restrictions on use.
- Data quality – the general reliability and quality, stated accuracy levels, level of completeness and consistency, and the source data.
- Data organization – the data model used to create and encode the information.

- Spatial referencing – the coordinate system used to georeference the data (if any), geographic datums used and any transformation parameters (relative to global standards) which may be specific to the geographic region.
- Attribute information – what kind of descriptive attribute system is used.
- Distribution – where the data were created, formats and other media types available, online availability, restrictions and costs.
- Metadata referencing – when, where and by whom the data (and metadata) were compiled.

Metadata have an important part to play in a world where digital analysis is commonplace and digital data are growing in volume all the time. The catch is that although metadata are very useful in increasing understanding about and tracking the provenance of data, they cannot prove the quality or the trustworthiness of the data.

## 11.5 GIS and the Internet

With easy-to-use Web browsers, GIS on the Internet provides a much more dynamic tool than a static map display. Web-enabled GIS brings interactive

query capabilities to a much wider audience. It allows online data commerce and the retrieval of data and specialized services from remote servers (such as that of the Environmental Systems Research Institute, Inc. (ESRI), the Geography Network ([www.geographynetwork.com](http://www.geographynetwork.com))).

The development of languages like GML, the Geography Markup Language (an extension of XML), VRML (superseded by X3D) and KML (Keyhole Markup Language) also make GIS far more accessible to the general, computer-literate public. KML, for instance, is a file structure for the storage and display of geographic data, such as points, lines, polygons and images, in Web browser applications such as Google Earth & Maps, MS Virtual Earth, ArcGIS Explorer, Adobe Photoshop and AutoCAD. It uses tags, attributes and nested elements and in the same way as standard HTML and XML files. KML files can be created in a simple text editor or in one of many script editing applications. More info can be found at [www.opengeospatial.org/standards/kml](http://www.opengeospatial.org/standards/kml). Such a facility is especially useful and relevant since it allows one to share geographical information with other users via the Internet, and, importantly, with other users who might not have access to proprietary GIS software, allowing them to see the data in a geographical or map/image context.

# 12

## Data Models and Structures

### 12.1 Introducing spatial data in representing geographic features

Data that describe a part of the Earth's surface or the features found on it could be described as *geographic* or '*spatial*' data. Such data include not only cartographic and scientific data, but also photographs, videos, land records, travel information, customer databases, property records, legal documents and so on. We also use the term 'geographic features' or simply 'features' in reference to objects, located at the surface of the Earth, whose positions have been measured and described. Features may be naturally occurring objects (rivers, vegetation), or anthropogenic constructions (roads, pipelines, buildings) and classifications (counties, land parcels, political divisions). Conventional cartographic maps represent the real world using collections of points, lines and areas, with additional textual information to describe the features. GIS constructs maps in a similar way but the features appearing on the map are stored as separate entities which have other intelligence stored with them as 'attributes'. It is also worth noting that the terms 'feature' and 'object' are commonly used interchangeably; here we have tried to stick to the term 'feature' when referring to a discrete entity, usually in vector form, but there are times when both terms are used for clarity.

### 12.2 How are spatial data different from other digital data?

There are four main aspects that qualify data as being spatial. Firstly, spatial data incorporate an explicit relationship between the geometric and attribute aspects of the information represented, so that both are always accessible. For instance, if some features are highlighted on a map display, the records containing the attributes of those features can also be highlighted (automatically) in the associated attribute table. If one or more of those features are edited in some way in the map, those changes are also automatically updated in the table, and vice versa. There is therefore a dynamic link between a feature's geometry and its attributes.

Secondly, spatial data are referenced to known locations on the Earth's surface, i.e. they are 'georeferenced'. To ensure that a location is accurately recorded, spatial data must conform to a coordinate system, a unit of measurement and a map projection. When spatial data are displayed, they have a specific scale just like on an analogue map, but in GIS this scale can be modified.

Spatial data also tend to be categorized according to the type of features they represent, i.e. they are sometimes described as being 'feature based'. For example, area features are stored separately from linear or point features and, in general, cannot coexist in the same file structure.

Lastly, spatial data are often organized into different ‘thematic’ layers, one for each set of features or phenomena being recorded. For instance, streams, roads, rail and land use would be stored as separate ‘layers’, rather than as one large file. In the same way, within each ‘theme’ there may be sub-themes of the same feature type which can be usefully grouped together, such as different classes of roads and railways, all of which are linear features that belong to the group ‘transport’. This also makes data management, manipulation and analysis rather more effective.

### 12.3 Attributes and measurement scales

Descriptive attributes can also be described as being spatial or non-spatial, though the difference between them may be subtle and ambiguous. The nature of the information stored, or rather the scale of measurement to which they belong, dictates what kind of processing or analysis can be performed with them. *Measurement scales* describe how values are assigned to features and objects represented in GIS. The type of scale chosen is dictated by the intended use of the recorded information. There are five scales commonly used in GIS, namely nominal, ordinal, interval, ratio and cyclic, and these are summarized, along with the numerical operators appropriate to each case, in Table 12.1.

*Nominal or categorical scales* include numerical values used to represent real-world objects or qualitative descriptions. They can be used as ‘pointers’ to other descriptive (textual) information held in attribute tables. *Ordinal measures* involve values ranked or ordered according to a relative scale and which generally have unequal intervals. Greater than or less than operators are therefore useful but addition, subtraction, multiplication and division are not appropriate. One example is multi-element geochemical data where element concentrations are given on a percentile scale and the intervals between classes are not constant but arbitrary. *Interval measures* are used to denote quantities like distances or ranges but in this case the intervals between the values are based on equal or regular units. There is, however, no true zero on an interval scale because the position of zero depends on the units of the quantity being described. Temperature scales are a good example because the position of zero temperature depends on the units of measurement, Fahrenheit or Celsius. *Ratio measures* are similar to interval scales and are often used for distances or quantities but the zero value represents absolute zero, regardless of the units. *Cyclic measures* are a special case describing quantities which are measured on regular scales but which are circular or cyclic in their framework, such as aspect or azimuth directions of slopes, or flow directions, both of which are angular measures made with respect to north. Appropriate operators are then any or all of the previously mentioned arithmetic and average operators.

**Table 12.1** Measurement scales: methods for describing and operating on thematic information

Scale	Operators	Examples
Nominal	=, ≠ and mode	Categorical (class) identifiers (e.g. 5 = forest, 4 = pasture, 9 = urban)
Ordinal	<, ≤, ≥, > and median	Sequences of natural order, for example 1, 2, 3, 4
Interval	+, −, ×, ÷ and mean	Ranges between, and sections along, distance measures, for example temperature scales
Ratio	All the above	Distance measures, and subdivisions thereof, along lines and routes
Cyclic	All the above	Special measures, for example 360° bearings (azimuth), flow directions

## 12.4 Fundamental data structures

There are two basic types of structures used to represent the features or objects, namely *raster* and *vector* data, and as a consequence of this split, there are different types of GIS software, and different types of analysis, which have been designed in such a way as to be effective with one or the other type.

Rasters, images or grids consist of a regular array of digital numbers or DNs, representing picture elements or pixels (as explained more thoroughly in Chapter 1) which are usually square. The basic unit of such data is the pixel, or grid cell, such that points, lines and areas are represented in raster form as individual or groups of pixels.

Vector or discrete data store the geometric form and location of a particular feature, along with its attribute information describing what the feature represents. Vector data typically resemble cartographic data.

*Points* and *pixels* represent discrete geographic features of no or limited area, or which are too small to be depicted in any other way, such as well locations, geochemical sample points, towns or topographic spot heights. *Lines* are linear features consisting of connected positions which do not in themselves represent area, such as roads, rivers, railways or elevation contours. *Areas* are closed features that represent the shape, area and location of homogeneous features such as countries, land parcels, buildings, rock types or land-use categories. A *surface* describes a variable which has a value for every position within the extent of the dataset, such as elevation or rainfall intensity, and implies data of a ‘continuous’ nature. Surfaces are typically represented on conventional cartographic maps as a series of isolines or contours; within GIS there are other possibilities. Deciding how these features should be stored in the database, and represented on the map, depends on the nature of that information and the work it will be required to do.

The two most basic components of GIS are therefore the pixel and the point. Every other, more complex, structure in GIS stems from, and depends on, one or other of these two basic structures. GIS operations and spatial analysis can be performed on either type of data, but that analysis will be performed

slightly differently as a result of this difference. We will now describe these structures in turn.

## 12.5 Raster data

As described more fully in Chapter 1, an image is a graphic representation or description of an object that is typically produced by an optical or electronic device. Some common examples of image data include remotely sensed (satellite or airborne survey data), scanned data and digital photographs.

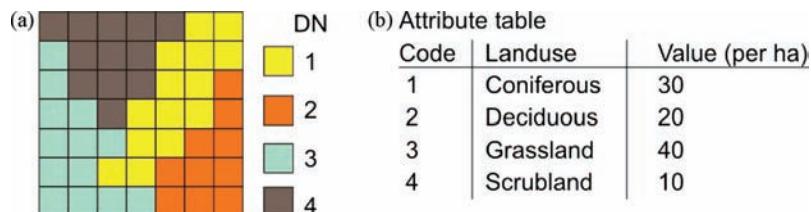
Raster data represent a regular grid or array of digital numbers, or pixels, where each has a value depending on how the image was captured and what it represents. For example, if the image is a remotely sensed satellite image, each pixel DN represents reflectance from a finite portion of the Earth’s surface; or, in the case of a scanned document, each pixel represents a brightness value associated with a particular point on the original document.

One important aspect of the raster data structure is that no additional attribute information is stored about the features it shows. Each raster grid therefore represents the variance of one attribute, such as reflectance or emittance from the Earth’s surface as in the case of remotely sensed data, so that a separate raster is needed to represent each attribute that is required. This means that image databases require considerably more disk space than their vector counterparts. In a large project this could become a serious consideration. Any additional attribute information must be stored in a separate table which can then be linked or related to that raster. For example, a raster in which integer values between 1 and 4 are stored, representing land-use categories, is linked with an attribute table which relates these values to other descriptive textual information, such as is illustrated in Figure 12.1.

If, for the same area, representation of other variables such as soil type or rainfall intensity is required, then these would have to be stored as separate raster images.

### 12.5.1 Data quantization and storage

The range of values that can be stored by image pixels depends on the quantization level of the



**Figure 12.1** A simple raster land-use map with other attributes stored in an associated table

data, i.e. the number of binary bits used to store the data. The more the number of bits, the greater the range of possible values. For example, if 1 bit data are used, the number of unique values that can be expressed is  $2^1$ , or 2. With 8 bits,  $2^8$  or 256 unique values can be expressed; with 16 bits, that number is  $2^{16}$  or 65 536 and so on. The most common image data quantization formats are 8 bit and 16 bit. Raster data can also be stored in ASCII format but this is by far the least efficient way of storing digital numbers and the file sizes get very large indeed. The binary quantization level selected depends partly on the type of data being represented and what it is used for. For instance, remotely sensed data are normally stored as 8 bit integer data, allowing 256 grey levels of image information to be stored, so that a three-band colour composite image provides three times 255 levels (8 bits) and hence 24 bit colour is produced. Digital elevation data, on the other hand, may well represent height values which are in excess of 256 m above sea level. The use of 8 bit data would not allow this, so elevation data are usually stored as 16 bit data, as integers or real numbers (floating-point values).

Image data can be organized in a number of ways and a number of standard formats are in common use. In many cases, the image data file contains a header record that stores information about the image, such as the number of rows and columns in the image, the number of bits per pixel and the georeferencing information. Following the image header is the actual pixel data for the image. Some formats contain only a single raster image, while others contain multiple images or bands. The file extension used usually reveals the method of storage of the image data, for example band interleaved by line has the extension *.bil*, band interleaved by pixel *.bip* and band sequential *.bsq*.

The main issue in connection with raster data storage is the disk space potentially required. With increasing availability of data, increasingly high resolution and increasing speed and computing power, so the capacity (and desire) to process large data volumes grows. In parallel with this has been the need to develop better methods of storage and compression. The goal of raster compression is then to reduce the amount of disk space consumed by the data file while retaining the maximum data quality. Newly developed image compression methods include *wavelet transforms*, which are produced using an algorithm based on multi-resolution analysis. Such methods are much less ‘lossy’ than block-based (discrete cosine transformation) compression techniques such as JPEG. The advantage of wavelet compression is that it analyses the whole image and allows high compression ratios while maintaining good image quality and without distorting image histograms. These techniques are essential if large volumes of raw data are to be served via the Internet/intranets. The most well known include the following and useful White Papers can be found on the OGC and relevant company websites:

- JPEG2000, a new standard in image coding, uses bi-orthogonal wavelet compression. It produces better compression ratios than its predecessor, JPEG, and is almost ‘lossless’.
- ECW or Enhanced Compressed Wavelet (copyright ER Mapping Ltd) uses multiple scale and resolution levels to compress imagery, while maintaining a level of quality close to the original image. Compression ratios of between 10 : 1 to 20 : 1 (panchromatic images) and 30 : 1 to 50 : 1 (colour images) are possible.
- MrSID or Multi-resolution Seamless Image Database (developed by Los Alamos National Laboratory and now marketed by LizardTech

Corporation) uses an algorithm very similar to ECW, with similar compression ratios and with similarly ‘lossless’ results, for serving over the Internet.

### 12.5.2 Spatial variability

The raster data model can represent discrete point, line and area features but is limited by the size of the pixel and by its regular grid-form nature. A point’s value would be assigned to and represented by the nearest pixel; similarly, a linear feature would be represented by a series of connected pixels; and an area would be shown as a group of connected pixels that most closely resembles the shape of that area, as illustrated in Figure 12.2.

The kind of information to be stored should be considered before choosing a model for storing it. Clearly points, lines and areas can be represented but since the raster grid is by definition contiguous, i.e. every possible position within the area of interest must be represented by a pixel value, we are forced to store values for all the areas not represented by significant features. So in the case of the land-use map in Figure 12.1, the information of value occurs at the boundaries between the classes. The distribution of scrubland for instance is such that it covers the area taken by 12 pixels, thus its value (4 in this case) is stored 12 times. There is no further variation within this area and a similar situation exists for the other three classes. Such information can be considered to be of low spatial variability and, if represented by a raster, means that many duplicate and so insignificant values are stored in addition to the important ones where the

class boundaries occur; this constitutes data redundancy and represents a waste of disk space. In such cases, it would be better to choose the vector model of storage.

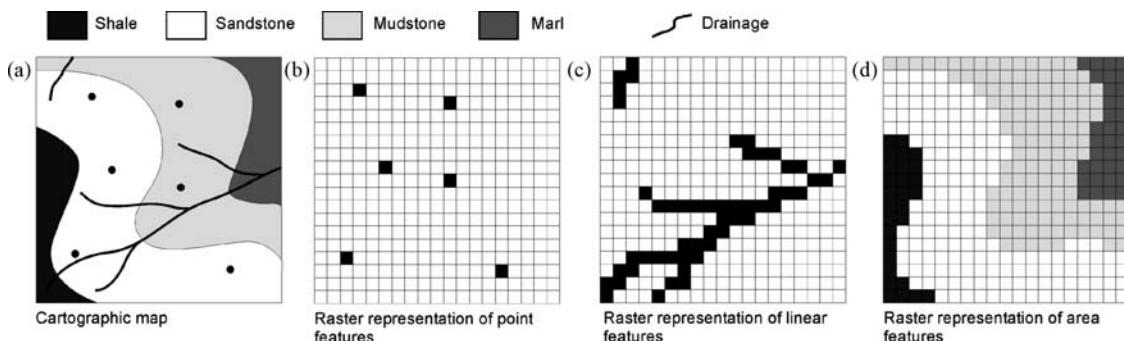
Since the raster model enforces the storage of every pixel DN, it is therefore most appropriately used where the data can be considered of relatively high spatial variability; i.e., where every pixel has a unique and significant value. This applies to satellite imagery or photographs, or where the objective of our analysis means that we are interested in the spatial variability of the attribute or phenomenon across an area, not just in discrete classes.

There may, however, be times during spatial analysis, for instance, when we need to convert such maps and images of low spatial variability into raster form. Usually such files are intermediary and they are normally automatically removed at the end of the analysis.

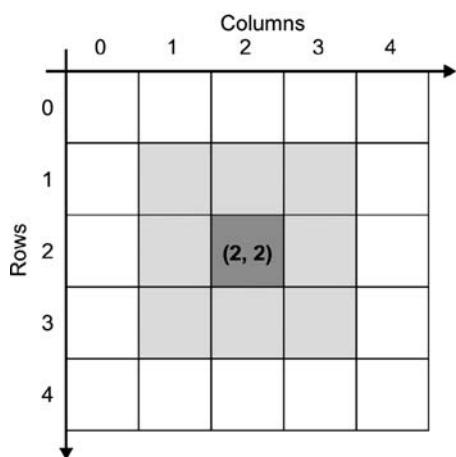
### 12.5.3 Representing spatial relationships

Because the raster data model is a regular grid, spatial relationships between pixels are implicit in the data structure since there can be no gaps or holes in the grid. Any further explicit description of spatial relationships is unnecessary.

Each raster is referenced at the top-left corner; its location is denoted by its row and column position and is usually given as 0, 0. All other pixels are then identified by their position in the grid relative to the top left. Each pixel in the raster grid has eight immediate neighbours



**Figure 12.2** Raster representation of cartographic features (point, line and area)



**Figure 12.3** The organization of a simple raster grid and its row and column reference system

(except those on the outside edges). In Figure 12.3, the cell at position 2, 2 is located at 3 pixels along the  $x$  axis and 3 down on the  $y$  axis. Finding any one of eight neighbours simply requires adding or subtracting from the  $x$  and/or  $y$  pixel locations. For example, the value immediately to the left of (2, 2) is (2 - 1, 2 or 1, 2).

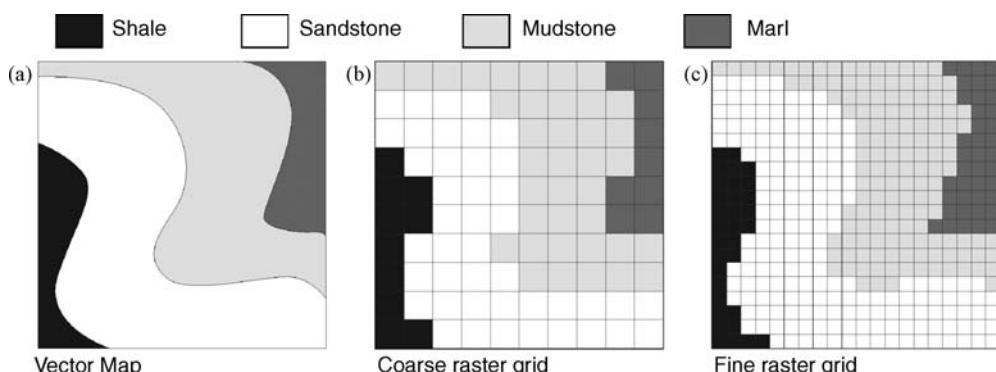
Since the spatial relationships are implicit, the whole the raster is effectively georeferenced by this system; specifying the real-world location of the 0, 0 reference point and the cell size in real-world distances enables rapid calculation of the real-world locations of all other positions in the grid. The upper left pixel being used as the reference point for ‘raster space’ is in contrast to ‘map space’

where the lower left corner is the geographical coordinate origin; this difference has an effect on the way raster images are georeferenced (as described in Chapter 13).

Another benefit of implicit spatial relationships is that spatial operations are readily facilitated. Processing can be carried out on each individual pixel value in isolation, between corresponding pixel positions in other layers, between a pixel value and its neighbour’s values or between a zone of pixel values and zones on other layers. This is discussed in more detail in Chapter 14.

#### 12.5.4 The effect of resolution

The accuracy of a map depends on the scale of that map. In the raster model the resolution, scale and hence accuracy depend on the real-world area represented by each pixel or grid cell. The larger the area represented, the lower the resolution of the data. The smaller the area covered by the pixel, the greater the resolution and the more accurately and precisely the features are represented. This is demonstrated in Figure 12.4, where the boundaries of the lithological units in the geological map are most closely represented by the raster grid whose spatial resolution is the highest, i.e. whose pixels are the smallest (Figure 12.4c). The problem arises at the boundaries of the classes; where a boundary passes through a pixel, rules need to be applied to decided how that pixel will be encoded. This issue is discussed further later in this chapter.



**Figure 12.4** The effect of raster resolution on precision of presentation

The raster data model may at first seem unappealing, within GIS, because of its apparent spatial inaccuracy, i.e. the limitation of the sampling interval or spatial resolution, but it more than makes up for this in its convenience for analytical operations (being a 2D array of numbers). This is especially true for any operations involving surfaces or overlay operations, and of course with remotely sensed images. The pixel can be thought of as the limit beyond which the raster becomes discrete, and with computer power becoming ever greater we may have fewer concerns over the manageability of large, high-resolution raster files. Providing we maintain sufficient spatial resolution to describe adequately the phenomenon of interest, we should be able to minimize problems related to accuracy.

#### 12.5.5 Representing surfaces

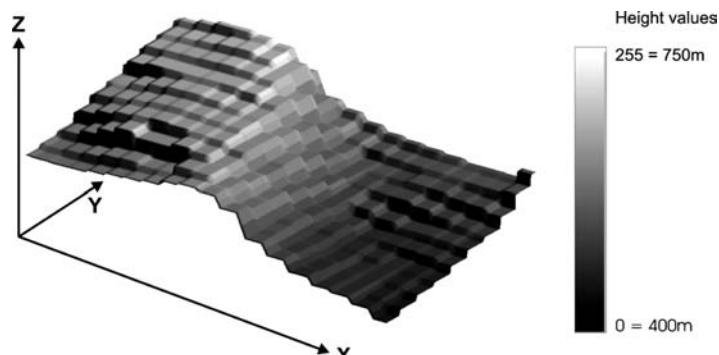
Rasters are ideal for representing surfaces since a value, such as elevation, is recorded in each pixel and the representation is therefore ‘continuously’ sampled across the area covered by the raster. Conceptually, we find it easiest to think of a surface, from which we can generate a perspective view, as being elevation or topography, but any raster can be used as a surface, with its pixel values (or DNs) being used to generate ‘height’ within 3D space, as illustrated in Figure 12.5. The structure and use of raster surfaces are dealt with in more detail in

Chapter 16. The input dataset representing the surface potentially contributes two pieces of information to this kind of perspective viewing. The first is the magnitude of the DN which gives the height, and the second is the way the surface appears or is encoded visually, i.e. the DN value is also mapped to colour in the display.

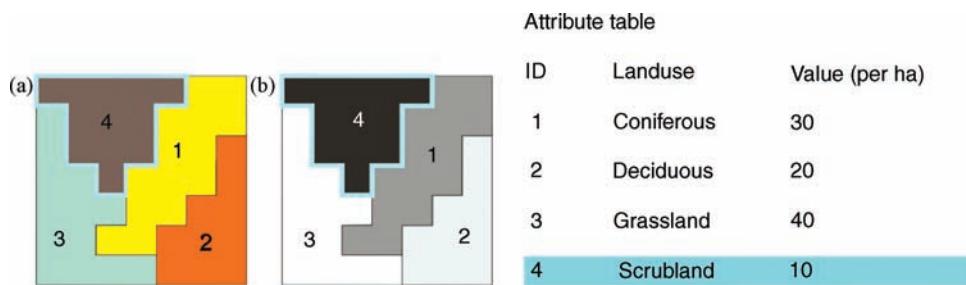
## 12.6 Vector data

In complete contrast, the vector model incorporates, discretely, both the geometry and location of geographic features and all the attribute information describing them. For convenience, the attribute information is generally viewed and accessed in tabular form. Each vector feature has a unique identifying code or number and then any number of numerical or textual attributes, such that the features can be symbolized according to any of those attributes. In Figure 12.6 a simple vector map is shown where the individual land parcels are coded to represent two stored attributes in Figure 12.6a and 12.6b respectively: land use (nominal) and land value (ordinal); the alphanumeric attributes used are shown on the right in Figure 12.6.

This association of geometry and tabular attribute is often referred to as a ‘georelational’ data structure and, in this way, there is no limit to the attribute information which can be stored or linked to a particular feature object. This represents one very



**Figure 12.5** A raster grid displayed, in perspective view, to illustrate the way that pixel values are used to denote height and to form a surface network. The pixel values can be colour coded on a relative scale, using a colour lookup table in which height values range from low to high, on a black to white scale in this case



**Figure 12.6** A simple vector file colour coded to show (a) land use (nominal) and (b) land value (ordinal), and its attributes, with feature number 4 highlighted in both the map and the attribute table

clear advantage over raster data's single attribute status. As with raster data, additional tables can also be associated with the primary vector feature attribute tables, to build further complexity into the information stored.

Tabular data represent a special form of vector data which can include almost any kind of data, whether or not they contain a geographic component; tabular data are not necessarily spatial in nature. A table whose information includes and is referenced by coordinates can be displayed directly on a map. The information which does not must be linked to other spatial data that do have coordinates before it can be displayed on a map.

Vector data therefore consist of a series of discrete features described by their coordinate positions rather than graphically or in any regularly structured way. The vector model could be thought of as the opposite of raster data in this respect, since it does not fill the space it occupies; not every conceivable location is represented, only those where some feature of interest exists. If we were to choose the vector model to represent some phenomenon that varies continuously and regularly across a region, such that the vector data necessarily become so densely populated as to resemble a raster grid, then we would probably have chosen the wrong data model for those data.

### 12.6.1 Representing logical relationships

While the term 'topography' describes the precise physical location and shape of geographical objects, the term 'topology' defines the logical relationships between those objects, so that data whose topologi-

cal relationships are defined can be considered intelligently structured. GIS can then determine where they are with respect to one another, as well as what they represent. The difference between these two concepts is illustrated in Figure 12.7.

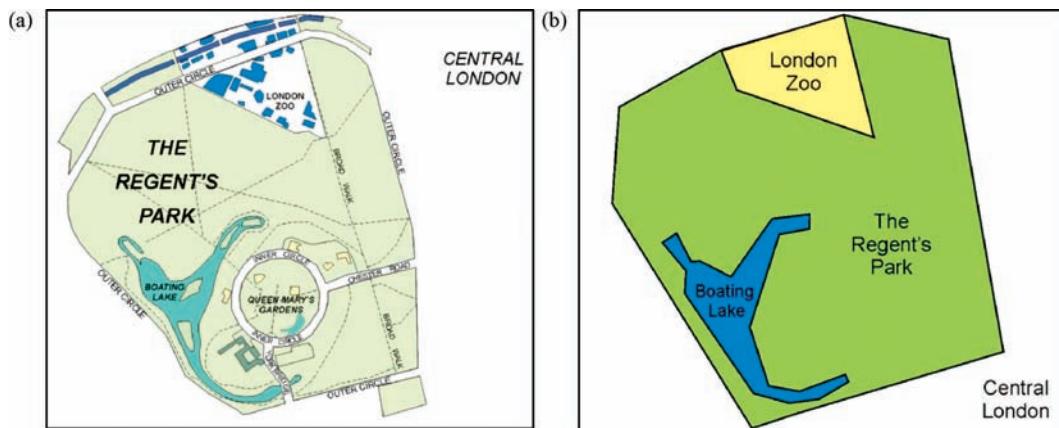
Looking at the topographic map in Figure 12.7, it is an easy for us to interpret the relationships – that Regent's Park completely encloses the Boating Lake and shares a boundary with the zoo – but it is not so for the computer. GIS can only understand these spatial relationships through topological constructs and rules.

There have been a number of vector data models developed over the past few decades, which support topological relationships to varying degrees, or not at all. The representation, or not, of topology dictates the level of functionality that is achievable using those data. These models include spaghetti (unstructured), vertex dictionary, dual independent map encoding (DIME) and arc-node (also known as POLYVRT). To understand the significance of topology it is useful to consider these models, from the simplest to the more complex.

#### 12.6.1.1 Unstructured or 'spaghetti' data

At the simplest level, each vector location is recorded as a single  $x, y$  coordinate pair, representing a *point*. Points are either isolated or joined to form lines, when each is then termed a *vertex*. Lines are then recorded as an ordered series of vertices, and areas are delimited by a series of ordered line segments which enclose that area.

The information describing the features is stored as a simple file listing the coordinate pairs of all the points comprising the feature and a unique identifying character for each feature. Three simple



**Figure 12.7** (a) A topographic map showing the Regent's Park area in London; and (b) a topological map showing the locations of London Zoo and the Boating Lake which lie inside Regent's Park in London

vector features (area, line and point respectively) and their file formats are shown in Figure 12.8. Each feature is identified by a unique code (A, B or C in this case) and by the number of vertices it contains, followed by the coordinate pairs defining all the constituent vertex locations. The order of the listed points comprising the area and line features is significant: they must be listed according to the sequence of connection. Notice also that the area feature is different from the line feature only in that the first and last listed points are identical, indicating that the feature is closed.

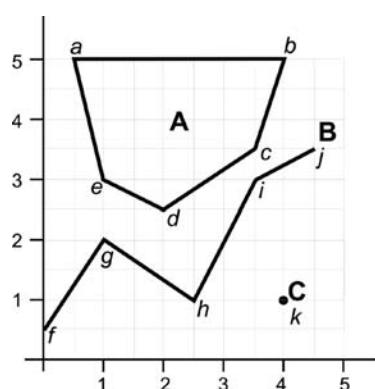
In this ‘spaghetti’ form, vector data are stored without relational information; there is no mechanism to describe how the features relate to one another, i.e. there is no topology. Any topological

relationships must be calculated from the coordinate positions of the features; this was considered inefficient and computationally costly.

The advantages of unstructured data are that their generation demands little effort (this is perhaps debatable) and that the plotting of large unstructured vector files is potentially faster than structured data. The disadvantages are that storage is inefficient, there may be redundant data and relationships between objects must be calculated each time they are required.

#### 12.6.1.2 Vertex dictionary

This structure is a minor modification of the ‘spaghetti’ model. It involves the use of two files to represent the map instead of one. Using the same



File stored as a simple list of coordinate pairs:

A, 6	(polygon identifier, number of vertices)
0.5, 5	(coordinates of first vertex (a))
4,	5
3.5, 3.5	
2,	2.5
1,	3
0.5,	3
B, 5	(line identifier, number of vertices)
0,	0.5
1,	2
2.5,	1
3.5,	3
4.5,	3.5
C, 1	(point identifier, number of vertices)
4,	1
	(coordinates of vertex (k))

**Figure 12.8** Three simple vector features stored as unstructured or ‘spaghetti’ data

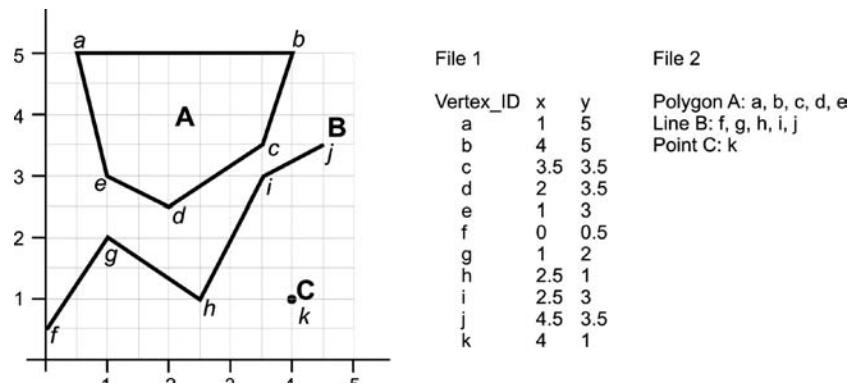


Figure 12.9 Vertex dictionary model

map shown in Figure 12.9 as an example, in the first file the coordinate pairs of all the vertices are stored as a list, with each pair being identified by a unique ID number. The second file stores the information describing which coordinate pairs are used by each feature. This prevents duplication, since each coordinate pair is stored only once, but it does not allow any facility to store the relationships between the features, i.e. topology is still not supported.

#### 12.6.1.3 Dual independent map encoding

The DIME structure was developed by the US Bureau of the Census for managing its population databases; both street addresses and UTM coordinates were assigned to each entity in the database. Here again, additional files (tables) are used to describe how the coordinate pairs are accessed and used (as shown in Figure 12.10). This time the vertices and arcs are given unique codes and each

arc is assigned a directional identifier to denote the direction in which the feature is constructed (the *from* and *to* vertices). In this way some topological functionality is supported, through the connectivity of features. Arcs which form polygons are listed in a further file.

The US Bureau of the Census later developed this and released the TIGER/Line (or Topologically Integrated Geographic Encoding and Referencing) format in the 1990s. TIGER incorporates a higher block level, thus a further level of hierarchy (with unique block ID numbers), to add complexity.

#### 12.6.1.4 Arc-node structure or POLYVRT (*POLY*gon *conVeRTer*)

Here vector entities are stored separately but are linked using pointers. A further concept of chains is also added, in which chains form collections of line segments with directional information.

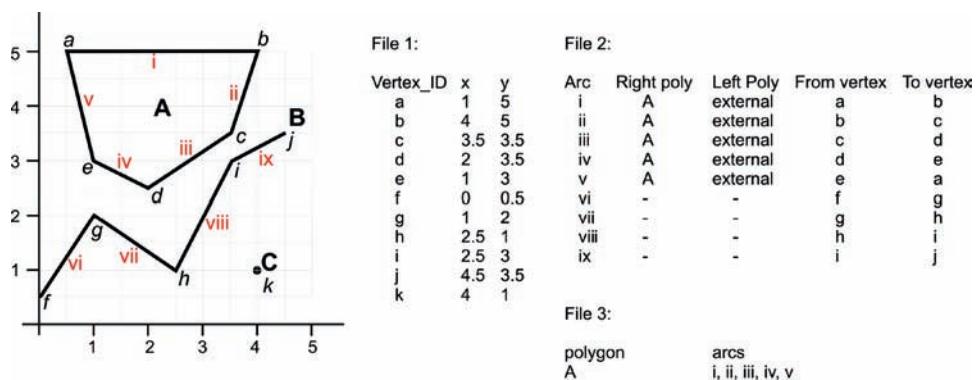


Figure 12.10 The DIME data structure

When capturing the boundaries of two adjacent area features from a map, the double tracing of the common boundary between them is inefficient and can lead to errors (either gaps or overlaps). The same is true for a point shared by a number of lines, because the point may be stored many times. A more efficient model for the storage of vector data, and one which supports topological relationships, is the ‘arc-node’ data structure.

We also have a potential problem in storing polygons that have holes or islands in them. A polygon that exists inside another polygon is often termed an ‘island polygon’ (or hole) but if any of the previously described data models are used, we have no way of conveying that one lies inside the other, they merely have location in common. There are ways to get around this with other models if, for whatever reason, the data must remain unstructured. If a value field is added to the polygon attribute table which denotes some level of priority, then this value can be used to control how the file is drawn or plotted on the screen or display; polygons with low-priority attributes would be drawn first and those with high priority would be drawn last. In this way the smaller, high-priority island polygons could be made to draw on top of the outer polygon that encloses them. If the polygon file in question contains lithological outcrop patterns in a geological map, for example, this task could get rather complicated with many levels of priority required to ensure the polygons are drawn in the correct sequence. Not an ideal solution!

A further level of complexity is therefore required to define properly such complex relationships and so some more definitions would be helpful at this point. A sequence of vertices forms a line where the first and last vertices may be referred to as ‘start’ and ‘end’ vertices, and these have special significance and confer the direction of capture. An *arc* is a line which, when linked with other arcs, forms a *polygon*. Arcs may be referred to as *edges* and sometimes as *chains*. A point where arcs terminate or connect is described as a *node*. Polygons are formed from an ordered sequence of arcs and may be termed ‘simple’ or ‘complex’ depending on their relationship to other polygons. Where one polygon completely encloses another, the outer polygon is described as being complex. The outer polygon’s boundary with the inner, island polygon may be

referred to as an *inner ring* (of which there may be more than one) and its other, outer boundary is called its *outer ring*. So one or more arcs form a *ring*.

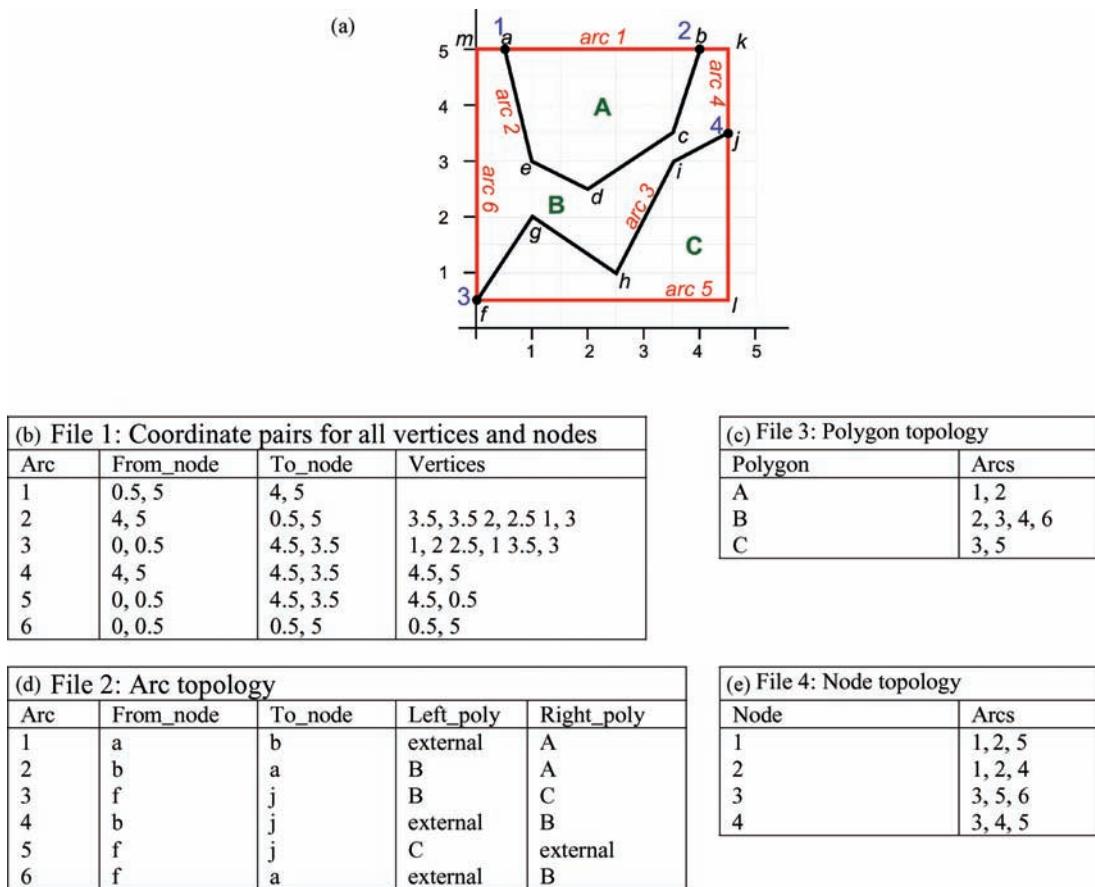
This scheme is created from the moment of capture, when the features are ‘digitized’. Each arc is digitized between nodes, in a consistent direction; it has a start and end node, and is given an identifying number. At the same time, further attribute information is entered which describes the identity of the polygon features that exist to the left and to the right of the arc being captured. When all the arcs are captured and identified in this way, the topological relationships between them are calculated and polygons are constructed. Usually some minor mistakes are made along the way but these can then be corrected and the topological construction process repeated until the polygon map is complete. The same principle is then applied to line maps, for the construction of networks for example.

A scheme of tables is used to describe how each of the levels of information relates to another. The coordinate pairs describing each arc are stored, including its *from* and *to* nodes, and any other vertices in between. Another table describes the arc topology, i.e. which polygon lies to the left or to the right of each arc (on the basis of the stored direction of each arc). A third table describes the polygon topology, listing the arcs that comprise each polygon. The last table describes the node topology, which lists the arcs that are constructed from each node.

For instance, consider the example used before (and shown in Figure 12.11), this time slightly modified by the addition of lines forming a boundary around the polygon and line, forming new polygons (three in total, labelled A, B and C). Nodes are created where more than two lines intersect, arcs are created between the nodes, with vertices providing the shape.

#### 12.6.1.5 New Structures

Since the development of POLYVRT, and in the context of superior computing power, a new generation of vector models has been formed which do not demand the rigorous construction of topologically correct data. The ESRI *shapefile* is a good example of this. In such cases, topological relationships are computed in memory, ‘on-the-fly’. In terms of the formats described here, the ESRI shapefile lies somewhere between ‘spaghetti’ and



**Figure 12.11** A simple map to illustrate the concept of arc-node structure

true topological structures. Thus the shapefile concept enforces the organization of data into separate types of shapefiles according to feature class type (e.g. point, polyline, polygon, pointZ, pointM, polylineZ, etc.; in all there are 14 possible types), by grouping into themed feature datasets, and by the use of complex file structures. The shapefile actually consists of a minimum of three files (.shp, .shx and .dbf) to store coordinates, geometry and attributes, with a further file (.prj) to describe map projection. This hybrid vector format provides some freedom from the requirement to capture rigorously and construct topologically correct data, but does not replace it. Generally speaking, vector models tend to be application specific: different application areas tend to have different demands on the data and so tend to adopt slightly different formats as standard.

Many GISs employ a relational database management system (DBMS) to connect the attribute information to the geometric information. More recently, object-oriented databases have been developed. These allow discrete objects to belong to discrete classes, and these may be given unique characteristics. Most modern GISs are hybrids of these in which the GIS functionality is closely integrated with the management system. Such systems allow vector and raster data to be managed and used together. The object-oriented type also integrates a spatial query language to extend the hybrid model; the ArcGIS geodatabase is an example of this type of structure. Both ESRI coverages and geodatabases are relational structures storing vector data, allowing geometry to be shared between attributes and vice versa. A *coverage* consists of a database directory holding the vector feature data

and its attributes as a series of files of specific names and extensions, according to the type of coverage. Both types use rules to validate the integrity of data and to derive topological relationships between features.

The use of topological relationships, however they are defined, has several clear advantages, including the more efficient storage of data, enabling large datasets to be processed quickly. One important advantage is that analytical functions are facilitated through three major topological concepts: *connectivity*, *area definition* and *contiguity*.

#### 12.6.1.6 Connectivity

Connectivity allows the identification of a pathway between two locations, between your home and the airport, along a bus, rail and/or underground route, for instance. Using the arc-node data structure, a route along an arc will be defined by two end points, the start or *from-node* and the finish or *to-node*. Network connectivity is then provided by an arc-node list that identifies which nodes will be used as the *from* and *to* positions along an arc. All connected arcs are then identified by searching for node numbers in common. In Figure 12.11, for example, it is possible to determine that arcs 1, 2 and 6 all intersect because they share node 1. GIS can then determine that it is possible to travel along arc 1 and turn onto arc 2 because they meet at node 1.

#### 12.6.1.7 Area definition

This is the concept by which it is determined that the Boating Lake lies completely within Regent's Park, i.e. it represents an island polygon inside it, as shown in Figure 12.7.

#### 12.6.1.8 Contiguity or adjacency

Contiguity, a related concept, allows the determination of adjacency between features. Two features can be considered adjacent if they share a boundary. Hence, the polygon representing London Zoo can be considered adjacent to Regent's Park as shown in Figure 12.8.

Remembering that the *from-node* and *to-node* define an arc's direction, so that the polygons on its left and right sides must also be known, *left-right topology* describes this relationship and therefore adjacency. In the arc-node data structure shown in Figure 12.11, polygon B lies to the left of arc 2, and

polygon A lies to the right, so we know that polygons A and B are adjacent.

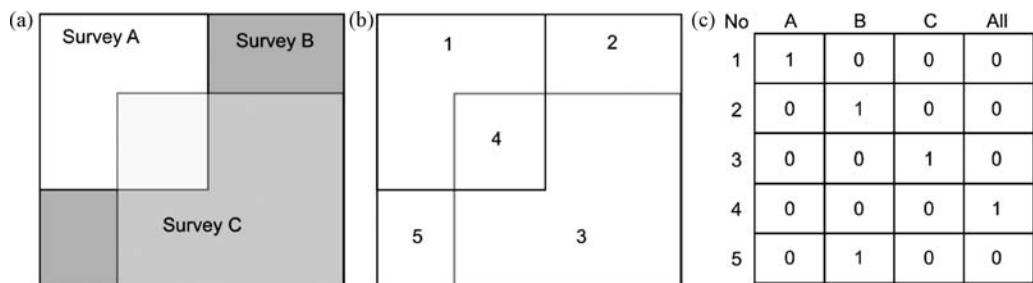
Notice the outer polygon which has no label here but is often called the *external or universe polygon*; it represents the world outside the area of interest and ensures that each arc always has a left and right side defined. So the arc joining points *a* and *b*, when captured in that order, has polygon A to its right and the universe polygon to its left. From this logic, entered at the time of data capture, topological entities are constructed from the arcs, to provide a complex and contiguous dataset.

#### 12.6.2 Extending the vector data model

Topology allows us to define areas and to model three types of association, namely connectivity, area definition and adjacency (or contiguity), but we may still need to add further complexity to the features we wish to describe. For instance, a feature may represent a composite of other features, so that a country could be modelled as the set of its counties, where the individual counties are also discrete and possibly geographically disparate features. Alternatively, a feature may change with time, and the historical tracking of the changes may be significant. For instance, a parcel of land might be subdivided and managed separately but the original shape, size and attribute information may also need to be retained. Other examples include naturally overlapping features of the same type, such as the territories or habitats of several species, or the marketing catchments of competing supermarkets, or surveys conducted in different years as part of an exploration program (as illustrated in Figure 12.12).

The 'spaghetti' model permits such area subdivision and/or overlap but cannot describe the relationships between the features. Arc-node topology can allow overlaps only by creating a new feature representing the area of overlap, and can only describe a feature's relationship with its subdivisions by recording that information in the attribute table.

Several new vector structures have been developed by ESRI and incorporated into its ArcGIS technology. These support and enable complex relationships and are referred to as regions, sections, routes and events.



**Figure 12.12** (a) Map of the boundaries of three survey areas, carried out at different times. Notice that the areas overlap in some areas; this is permitted in ‘spaghetti’ data but not in arc-node structures; (b) the same survey maps after topological enforcement to create mutually exclusive polygonal areas; (c) the attribute table necessary to link the newly created polygons (1 to 5) to the original survey extents (A, B and C). Modified after Bonham-Carter (2002)

### 12.6.2.1 Regions

A *region* consists of a loose association of related polygons and allows the description of the relationships between them. In the same way that a polygon consists of a series of arcs and vertices, a series of polygons form a region. The only difference in the structure is that, unlike the vertices listed in a polygon, the polygons comprising the region may be listed in any order. As with points, lines and polygons, each region has a unique identifier. The polygons representing features within the region are independent, they may overlap and they do not necessarily cover the entire area represented by the region. So the overlapping survey areas in Figure 12.12 could simply be associated within a survey region, as related but separate entities, without having to create a new, topologically correct polygon for the overlap area.

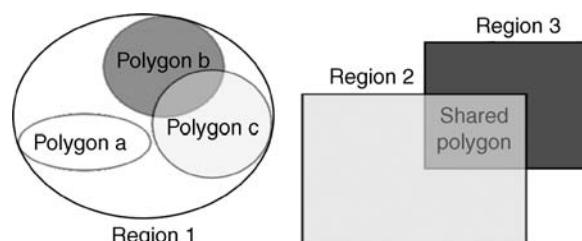
Normally, each feature would be represented by a separate polygon but the region structure allows for a single feature to consist of several polygons. For example, the islands around Great Britain might be

stored as independent and unconnected polygons but could belong to the collective region of the United Kingdom and be given collective properties accordingly.

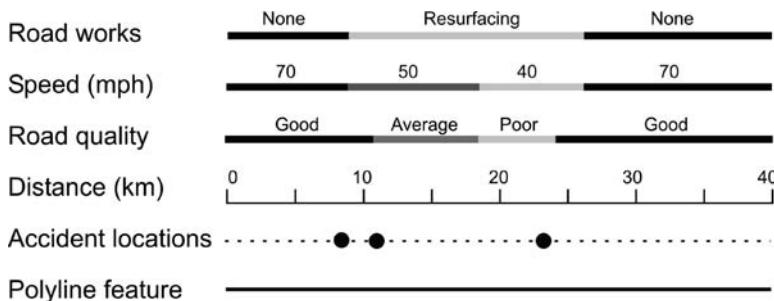
Constructing overlapping regions is rather similar to constructing polygons; where regions overlap, they share a polygon in the same way that polygons share an arc where they meet, as shown in Figure 12.13. The use of regions should assist data management since several different types of features may be integrated into a single structure while the original characteristics remain unchanged.

### 12.6.2.2 Linear referencing

Routes, sections and events can be considered together since they tend not to exist on their own, and together they constitute a system of *linear referencing* as it is termed in ESRI’s ArcGIS. The constructed route defines a new path along an existing linear feature or series of features, as illustrated in Figure 12.14. If we use the example of getting from your home to the airport, your



**Figure 12.13** Illustration of different types of region: associations of polygons and overlapping polygons which share a polygon



**Figure 12.14** Several routes representing different measures (linear and point events), created from and related to a pre-existing polyline feature representing, in this case, a road network

'airport' route would consist of a starting and finishing location and a unique path which follows existing roads, bus routes or railways but does not necessarily involve the entire lengths of those roads and railways. A *route* may be circular, beginning and ending in the same place. Routes may be disconnected, such as one that passes through a tunnel and so is not visible at the surface. A further piece of information necessary for the description of a route is the unit of measurement along the route. This could be almost any quantity and for the example of a journey the measure could be time or distance.

A *section* describes particular portions of a route, such as where road works are in progress on a motorway, where speed limits are in place on a road, or where a portion of a pipeline is currently undergoing maintenance. Again, starting and ending nodes of the section must be defined according to the particular measure along the route.

Similarly an *event* describes specific occurrences along a route, and events can be further subdivided into point and linear events. A *point event* describes the position of a point feature along a route, such as an accident on a section of motorway or a leak along a pipeline. The point event's position is described by a measure of, for instance, distance along the route. A *linear event* describes the extent of a linear feature along a route, such as speed restrictions along a motorway, and is rather similar in function to a section. A linear event is identified by measures denoting the positions where the event begins and ends along the route.

Route and event structures are of use in the description of application-specific entities such as seismic lines and shot-point positions, since

conventional vector structures cannot inherently describe the significance of discrete measurements along such structures. Along seismic lines the shot-points are the significant units of measurement but they are not necessarily regularly spaced or numbered along that line, so they do not necessarily denote distance along it or any predictable quantity. The use of routes and events becomes an elegant method of accommodating such intelligence within GIS since the route can be annotated with a measure which is independent of its inherent geometric properties.

### 12.6.3 Representing surfaces

The vector data model provides several options for surface representation: isolines (or contours), the triangulated irregular network, or TIN, and, although less commonly used, Thiessen polygons. Contours can only describe the surfaces from which they were generated and so do not readily facilitate the calculation of further surface parameters, such as slope angle, or aspect (the facing direction of that slope); both of these are important for any kind of 'terrain' or surface analysis. The techniques surrounding the calculation of contours are comprehensively covered in many other texts and so we will skirt around this issue here.

#### 12.6.3.1 The TIN surface model or tessellation

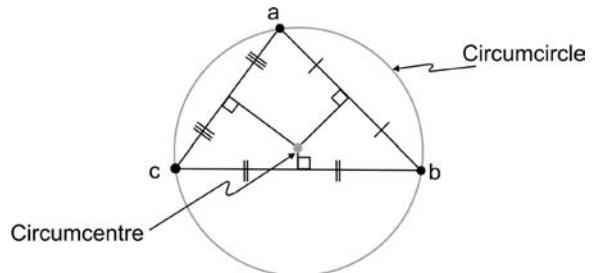
The TIN data model describes a 3D surface composed of a series of irregularly shaped and linked but non-overlapping triangles. The TIN is also sometimes referred to as the 'irregular triangular mesh'

or ‘irregular triangular surface model’. The points which define the triangles can occur at any location, hence the irregular shapes. This method of surface description differs from the raster model in three ways. Firstly, it is irregular in contrast with the regular spacing of the raster grid; secondly, the TIN allows the density of point spacing (and hence triangles) to be higher in areas of greater surface complexity (and requires fewer points in areas of low surface complexity); and lastly it also incorporates the topological relationships between the triangles.

The process of *Delaunay triangulation* (formulated by Boris Delaunay in 1934) is used to connect the input points to construct the triangular network. The triangles are constructed and arranged so that no point lies inside the *circumcircle* of any triangle (see Figure 12.15). Delaunay triangulation maximizes the smallest of the internal angles and so tends to produce ‘fat’ rather than ‘thin’ triangles.

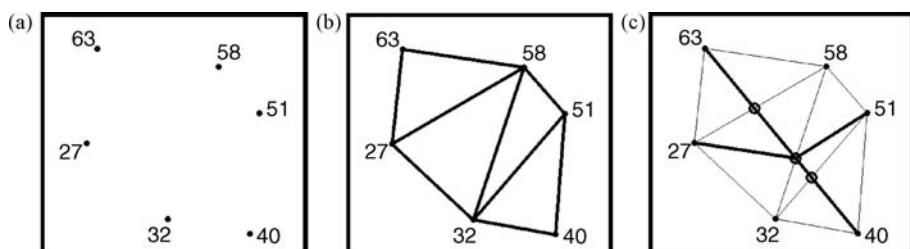
As with all other vector structures, the basic components of the TIN model are the points or nodes, and these can be any set of mass points with which are stored values other than  $x$ ,  $y$  and a unique identifying number, i.e. a  $z$  value in their attribute table (see Chapter 15 for more on  $z$  values). Nodes are connected to their nearest neighbours by edges, according to the Delaunay triangulation process. Left-right topology is associated with the edges to identify adjacent triangles. Triangles are constructed and break-lines can be incorporated to provide constraints on the surface.

The input mass points may be located anywhere. Of course the more carefully positioned they are, the more closely the model will represent the actual surface. TINs are sometimes generated from raster



**Figure 12.15** The Delaunay triangle constructed from three points by derivation of the circumcircle and circumcentre; the position of the latter is given by the intersection of the perpendicular bisectors from the three edges of the triangle

elevation models, in which case the points are located according to an algorithm that determines the sampling ratio necessary to describe the surface adequately. Well-placed mass points occur at the main changes in the shape of the surface, such as ridges, valley floors, or at the tops and bottoms of cliffs. By connecting points along a ridge or cliff, a break-line in the surface can be defined. By way of a simple example, for an original set of mass points (as shown in Figure 12.16a) the resultant constructed TIN is formed with the input point elevations becoming the TIN node elevations (Figure 12.16b). If this TIN is found to have undersampled (and so aliased) a known topographic complexity, such as a valley, a break-line can be included, such as a river. This then allows the generation of additional nodes at the intersection points with the existing triangles, and thereby further triangles are generated better to model the shape of the valley (Figure 12.16c).



**Figure 12.16** (a) Set of mass points, (b) the resulting TIN and (c) the new set of triangles and nodes formed by addition of a break-line to the TIN

Once the TIN is constructed, the elevation of any position on its surface can be interpolated using the  $x$ ,  $y$ ,  $z$  coordinates of the bounding triangle's vertices. The slope and aspect angles of each triangle are also calculated during TIN construction, since these are constant for each one.

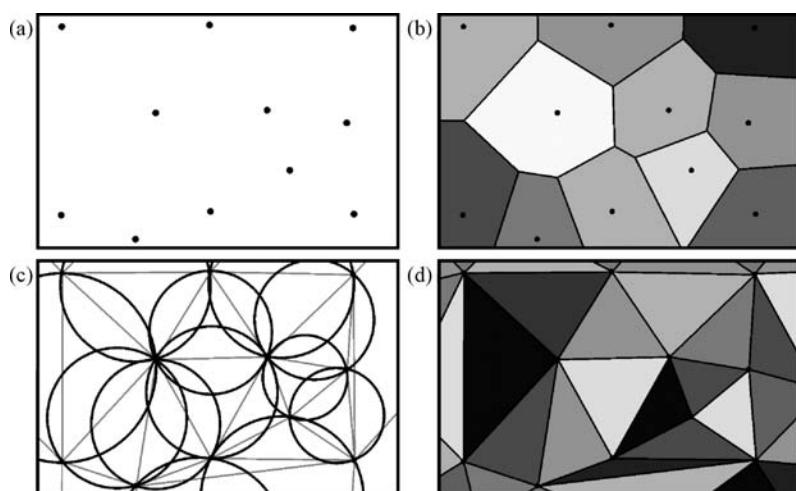
Another description of a surface can be given using Voronoi polygons (named after Georgy Voronoi). It can be said that the Delaunay triangulation of the set of points is equivalent to the dual graph, or is the topological and geometric dual, of the Voronoi tessellation (Whitney, 1932). The Voronoi tessellation consists of a set of lines, within a plane, that divide the plane into the area closest to each of a set of points; The Voronoi polygons are formed by lines that fall at exactly half the distance between the mass points and are perpendicular to the Delaunay triangle edges. Each Voronoi polygon is then assigned the  $z$  value of the point which lies at its centre. Once polygons are created, the neighbours of any point are defined as any other point whose polygon shares a boundary with that point. The relationship between the Voronoi polygons and the input points and Delaunay triangles is shown in Figure 12.17.

TINs allow rapid display and manipulation but have some limitations. The detail with which the surface morphology is represented depends on the number and density of the mass points and so

the number of triangles. So to represent a surface as well and as continuously as a raster grid, the point density would have to match or exceed the spatial resolution of the raster. Further, while TIN generation involves the automatic calculation of slope angle and aspect for each triangle, in the process of its generation, the calculation and representation of other surface morphological parameters, such as curvature, are rather more complex and generally best left in the realm of the raster. The derivation of surface parameters is dealt with in Chapter 15.

## 12.7 Conversion between data models and structures

There are sometimes circumstances when conversion from raster to vector formats is necessary for display and/or analysis. Data may have been captured in raster form through scanning, for instance, but may be needed for analysis in vector form (e.g. elevation contours needed to generate a surface, from a scanned paper topographic map). Data may have been digitized in vector form but subsequently needed in raster form for input to some multi-criteria analysis. In such cases it is necessary to convert between models and some consideration is required as to the optimum method, according to the stored attributes or the final intended use of the



**Figure 12.17** (a) The input 'mass' points, (b) Voronoi polygons constructed from the input points, (c) circumcircles (black) constructed from the input points and Delaunay triangles, and (d) the final TIN surface

**Table 12.2** Summary of general conversions between feature types (points, lines and areas), in vector/raster form

Conversion type	To point/pixel	To line	To polygon/area
From point/pixel	Grid or lattice creation	Contouring, line scan conversion/filling	Building topology, TIN, Thiessen polygons/interpolation, dilation
From line	Vector intersection, line splitting	Generalizing, smoothing, thinning	Buffer generation, dilation
From polygon/area	Centre point derivation, vector intersection	Area collapse, skeletonization, erosion, thinning	Clipping, subdivision, merging

product. There are a number of processes which fall under this description and these are summarized in Table 12.2.

### 12.7.1 Vector to raster conversion (rasterization)

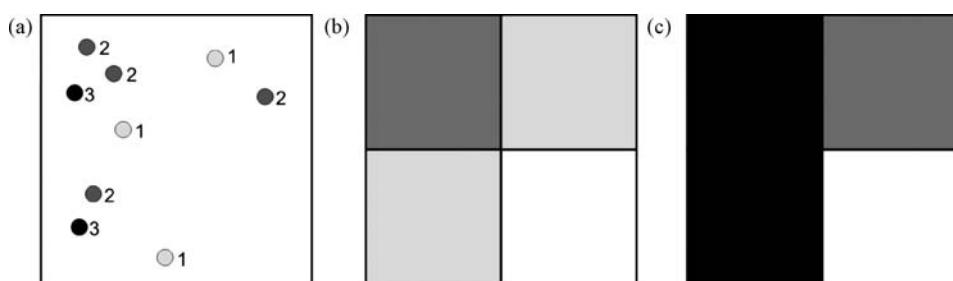
These processes begin with the identification of pixels that approximate significant points, and then pixels representing lines are found to connect those points. The locations of features are precisely defined within the vector coordinate space but the raster version can only approximate the original locations, so the level of approximation depends on the spatial resolution of the raster. The finer the resolution, the more closely the raster will represent the vector feature.

Many GIS programs require a blank raster grid as a starting point for these vector–raster conversions

where, for instance, every pixel value is 0 or has a null or no data value. During the conversion, any pixels that correspond to vector features are then ‘turned on’: their values are assigned a numerical value to represent the vector feature.

#### 12.7.1.1 Point to raster

For conversions between vector points and a discrete raster representation of the point data, there are several ways to assign a point’s value to each pixel (as shown in Figure 12.18). The first is to record the value of the unique identifier from each vector point. In this case, when more than one vector feature lies within the area of a single pixel, there is a further option to accept the value of either the first or the last point encountered since there may be more than one within the area of the pixel. Another option is to record a value representing merely the presence of a point or points. The third choice is to record the frequency of points found within a pixel. The fourth



**Figure 12.18** (a) Input vector point map (showing attribute values). (b) and (c) Two different resulting raster versions based on a most frequently occurring value rule (if there is no dominantly occurring value, then the lowest value is used) (b), and a highest priority class rule (where the attribute values 1–3 are used to denote increasing priority) (c)

is to record the sum of the unique identifying numbers of all vector points that fall with the area of the output pixel. The last is to record the highest priority value according to the range of values encountered.

Point, line and polygon features can be converted to a raster using either textual or numerical attribute values. Only numbers are stored in the raster file – numbers in a value range which dictates how the raster data are quantized, as byte or integer data for instance. So if text fields are needed to describe the information in the output raster, an attribute table must be used to relate each unique raster DN to its text descriptor. When pixels do not encounter a point, they are usually assigned a null (NoData) or zero value. The last is to record the highest priority value according to the range of values encountered.

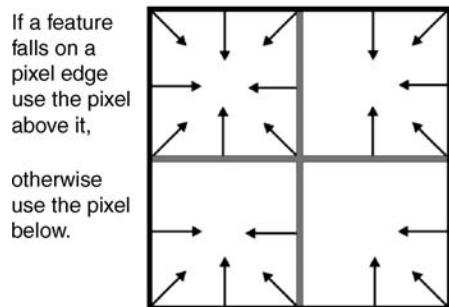
The above rules decide the value assigned to the pixel but further rules are required when points fall exactly on the boundary between pixels. These are used to determine which pixel will be assigned the appropriate point value. The scheme used within ESRI's ArcGIS is illustrated in Figure 12.19, in which a kind of kernel and associated logical rules provide consistency by selecting the edge and direction to which the value will be assigned.

Point to raster area conversions also include conversion to the continuous raster model. This category generally implies interpolation or gridding, of which there are many different types. These processes are dealt with in Chapter 15 rather than here.

#### 12.7.1.2 Polyline to raster

A typical line rasterizing algorithm first finds a set of pixels that approximate the locations of nodes. Then lines joining these nodes are approximated by adding new pixels from one node to the next one and

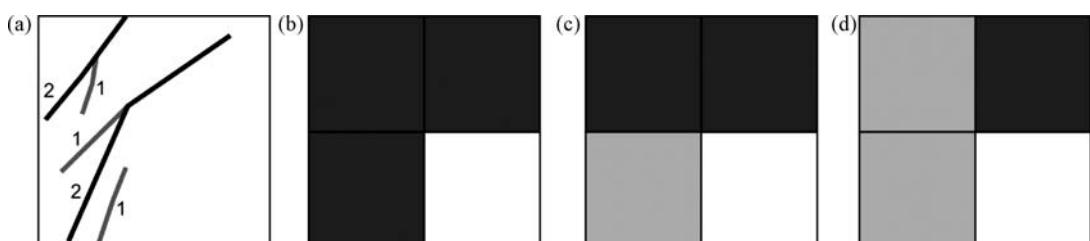
If there is a pixel to the left, use that.	Otherwise use the one to the right.
--	-------------------------------------



**Figure 12.19** Boundary inclusion rules applied when a feature falls exactly on a pixel boundary. Arrows indicate the directional assignment of attribute values. Modified after ESRI's ArcGIS online knowledge base

so on until the line is complete. As with points, the value assigned to each pixel when a line intersects it is determined by a series of rules. If intersected by more than one feature, the cell can be assigned the value of the first line it encounters, or merely the presence of a line (as with point conversions above), or of the line feature with the maximum length, or of the feature with the maximum combined length (if more than one feature with the same feature ID cross it), or of the feature that is given a higher priority feature ID (as shown in Figure 12.20). Again, pixels which are not intersected by a line are assigned a null or NoData value. Should the feature fall exactly on a pixel boundary, the same rules are applied to determine which pixel is assigned the line feature value, as illustrated in Figure 12.19.

The rasterizing process of a linear object initially produces a jagged line, of differing thickness along



**Figure 12.20** (a) Input vector line map. (b) to (d) Three different resulting raster versions based on a maximum length rule (or presence/absence rule) (b), maximum combined length rule (c) and a highest priority class rule (d) where the numbers indicate the priority attribute values

its length, and this effect is referred to as *aliasing*. This is visually unappealing and therefore undesirable but it can be corrected by anti-aliasing techniques such as smoothing. When rasterizing a line or arc the objective is to approximate its shape as closely as possible, but, of course, the spatial resolution of the output raster has a significant effect on this.

#### 12.7.1.3 Polygon to raster

The procedures used in rasterizing polygons are sometimes referred to as ‘polygon scan conversion’ algorithms. These processes begin with the establishment of pixel representations of points and lines that define the outline of the polygon. Once the outline is found, interior pixels are identified according to inclusion criteria; these determine which pixels that are close to the polygon’s edge should be included and which ones should be rejected. Then the pixels inside the polygon are assigned the polygon’s identifying or attribute value. This value will be found from the pixel that intersects the polygon centre.

The inclusion criteria in this process may be one of the following, whose effects are illustrated in Figure 12.21:

1. Central point rasterizing, where the pixel is assigned the value of the feature which lies at its centre.
2. Dominant unit or largest share rasterizing, where a pixel is assigned the value of the feature (or features) that occupies the largest proportion of that pixel.
3. Most significant class rasterizing, where priority can be given to a certain value or type of feature, such that if it is encountered anywhere within the area of a pixel, the pixel is assigned its value.

When viewed in detail (as in Figures 12.18, 12.20 and 20.21) it can be seen that the inclusion criteria have quite different effects on the form of the raster version of the input vector feature.

Again, if the polygon feature’s edge falls exactly on a pixel edge, special boundary rules are applied to determine which pixel is assigned the line feature value, as illustrated in Figure 12.19.

#### 12.7.2 Raster to vector conversion (vectorization)

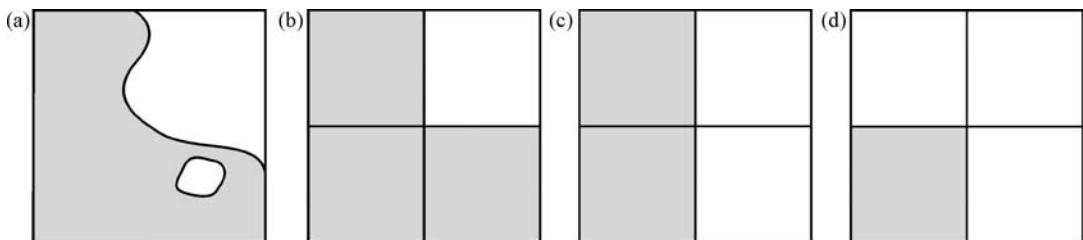
##### 12.7.2.1 Raster to point

All non-zero cells are considered points and will become vector points with their identifiers equal to the DN value of the pixel. The input image should contain zeros except for the cells that are converted to be points. The  $x$ ,  $y$  position of the point is determined by the output point coordinates of the pixel centroid.

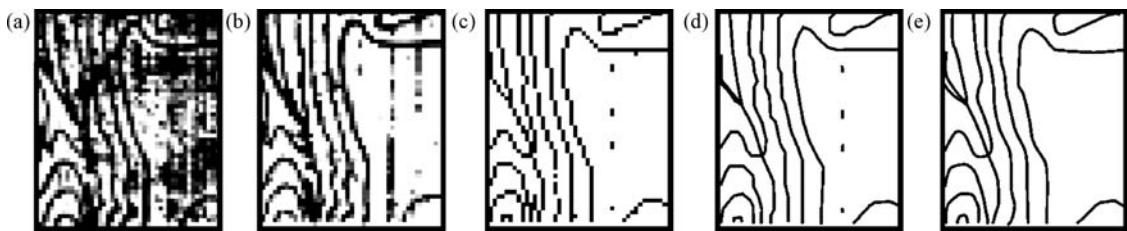
##### 12.7.2.2 Raster to polyline

This process essentially traces the positions of any non-zero or non-null raster pixels to produce a vector polyline feature, summarized in Figure 12.22. One general requirement is that all the other pixel values should be zero or a constant value. Not surprisingly, the input cell size dictates the precision with which the output vertices are located. The higher the spatial resolution of the input raster, the more precisely located the vertices will be. The procedure is not simple and generally involves several steps and is summarized as follows:

1. *Filling*: The image is first converted from a grey-scale to a binary raster (through reclassification



**Figure 12.21** (a) Input vector polygon map. (b) to (d) Three different resulting raster versions based on a dominant share rule (b), a central point rule (c) and a most significant class rule (d)



**Figure 12.22** Illustration of polygon scan vectorization procedures: (a) image improvement by conversion to binary (bi-level) image; (b) thinning or skeletonization process to reduce thickness of features to a single line of pixels; (c) vectorized lines representing complex areas; (d) collapsed but still including spurious line segments; and (e) the lines are smoothed or ‘generalized’ to correct the pixelated appearance and line segments removed

- or thresholding), then any gaps in the features are filled by *dilation*. The processes of filling and dilating are described in Chapter 14.
2. *Thinning*: The line features are then *thinned*, *skeletonized* or *eroded*, i.e. the edge pixels are removed in order to reduce the line features to an array or line of single but connected pixels. More about thinning and erosion appears in Chapter 14.
  3. *Vectorizing*: The vertices are then created and defined at the centroids of pixels representing nodes, that is where there is a change in orientation of the feature. The lines are produced from any connected chains of pixels that have identical DN value. The resultant lines pass through the pixel centres. During this process, many small and superfluous vertices are often created and these must be removed. The vectors produced may also be complex and represent area instead of a true linear feature.
  4. *Collapsing*: Complex features are then simplified by reducing the initial number of nodes, lines and polygons and, ideally, collapsing them to their centre lines. One commonly adopted method is that proposed by Douglas and Peucker (1974), which has subsequently been used and modified by many other authors.
  5. *Smoothing*: The previous steps tend to create a jagged, pixelated line, producing an appearance which is rather unattractive to the eye; the vector features are then smoothed or generalized, to smooth this appearance and to remove unnecessary vertices. This smoothing may be achieved by reducing the number of vertices, or using an averaging process (e.g. a three- or five-point moving average).

#### 12.7.2.3 Raster to polygon

This is the process of vectorizing areas or regions from a raster. Unless the raster areas are all entirely discrete and have no shared boundaries, it is likely that the result will be quite complex. Commonly, therefore, this process leads to the generation of both a line and a polygon file, in addition to a point file representing the centres of the output polygons

The polygon features are constructed from groups of connected pixels whose values are the same. The process begins by determining the intersection points of the area boundaries and then follows this by generating lines at either external pixel centroids or the boundaries. A background polygon is also generated, otherwise any isolated polygons produced will float in space. Again, such vectorization procedures from raster images are usually followed by a smoothing or generalization procedure, to correct the ‘pixelated’ appearance of the output vectors.

There are now a great many software suites available which provide a wealth of tools to perform these raster–vector conversions, some of which are proprietary and some ‘shareware’, such as MATLAB (MathWorks), AutoCAD (Autodesk), R2V (developed by Able Software Corp.), Illustrator, Freehand, etc.

## 12.8 Summary

Understanding the advantages and limitations of particular methods of representing data is key not only to effective storage and functionality but also to production of a reliable/accurate result. Knowing

how each data structure is used to describe particular objects, and what the limitations or drawbacks might be, is also useful.

The ability to describe topological relationships is one significant advantage of the vector model. The ability to describe (numerically) the relationships between objects is very powerful indeed and has no real equivalent in raster processing. Spatial contiguity, connectivity and adjacency have to be inferred through the implicit relationship of one pixel to another, but there is no inherent intelligence such as is enabled through topology. This may not present too much of a disadvantage in geoscientific analyses since we are often more interested in the relationship between different parameter values at one geographical position than in the relationship between objects which are geographically separated. In this sense, again it is the local or neighbourhood raster operations (discussed in Chapters 4 and 14) which gain our attention.

## Questions

12.1 Why is it important to understand the scales or levels of data measurement when observing and recording information?

- 12.2 What are ‘continuously sampled’ and ‘discrete’ data?
- 12.3 What are the advantages of using one structure over another?
- 12.4 What are the practical applications for each of these structures?
- 12.5 How should you decide on the most appropriate structure for a dataset?
- 12.6 Why does the *shapefile* not include topographical descriptions?
- 12.7 What are the differences between the topological vector model and the spaghetti vector model? What are the advantages and disadvantages of using each one?
- 12.8 How should more complex vector features (regions, routes and events) be organized?
- 12.9 What are the implications of spatial resolution on the raster representation of digital information and on accuracy?
- 12.10 What problems can arise during conversion between data models?
- 12.11 Are all these structures available in all software products?
- 12.12 What functionality do we need in a GIS software product to handle all these structures?
- 12.13 Do we need to invest in proprietary GIS software in order to work with geographic data?

# 13

## Defining a Coordinate Space

### 13.1 Introduction

Map projections and datums have been described very comprehensively by many other authors and we do not wish to repeat or compete with them, but this topic is an extremely important one in the understanding of GIS construction and functionality, and as such cannot be ignored here. We therefore prefer to overview the main principles, concentrating on the practical applications, and refer the reader to other more detailed texts.

To make GIS function, we must be able to assign ‘coordinates’ of time and location in a way that is generally understood. Calendar and temporal information can easily be included later as attributes or in metadata. Location is therefore the most important reference for data in GIS. Several terms are commonly used to denote the positioning of objects: georeference, geolocation, georegistration and geocoding. The main requirement for a georeference is that it is unique, to avoid any confusion. Hence the address, or georeference, of the Royal School of Mines, Prince Consort Road, London SW7 2AZ, United Kingdom, refers only to one building; no other in the world has this specific address. Georeferencing must also be persistent through time, again to avoid both confusion and expense.

Every georeference also has an implication of resolution, that is to a specific building, or collection of buildings, or a region. Many georeferencing

systems are unique only within a limited area or domain; for example, within a city or county. There may be several towns called Boston but only one is in Lincolnshire. Some georeferences are based on names and others on measurements. The latter are known as ‘metric’ georeferences and they include latitude/longitude and other kinds of regular coordinate systems; such metric coordinates are more useful to us as geoscientists. Some coordinate systems involve combinations of metric measures and textual information. For instance, the six digits of a UK National Grid reference repeat every 100 km so that additional information (letters are used in this case) is required to achieve country-wide uniqueness. Metric systems provide infinitely fine resolution to enable accurate measurements. So how are systems for describing an object’s location (and measurements associated with it) established? The answer requires a metaphorical step backwards and consideration of the shape of the Earth.

### 13.2 Datums and projections

The Earth is a 3D object, roughly oblatelike spherical in shape, and we need to represent that 3D shape in a 2D environment, on paper or on a computer screen. This is the reason for the existence of a multitude of map projections – since this cannot be done without distorting information, accurate measurements

become potentially ambiguous. To achieve this 2D representation, two things need to be approximated: the shape of the Earth and the transformations necessary to plot a location's position on the map.

### 13.2.1 Describing and measuring the Earth

The system of latitude and longitude is considered the most comprehensive and globally constant method of description and is often referred to as the geographic system of coordinates, or *geodetic system*, and it is the root for all other systems. It is based on the Earth's rotation about its centre of mass.

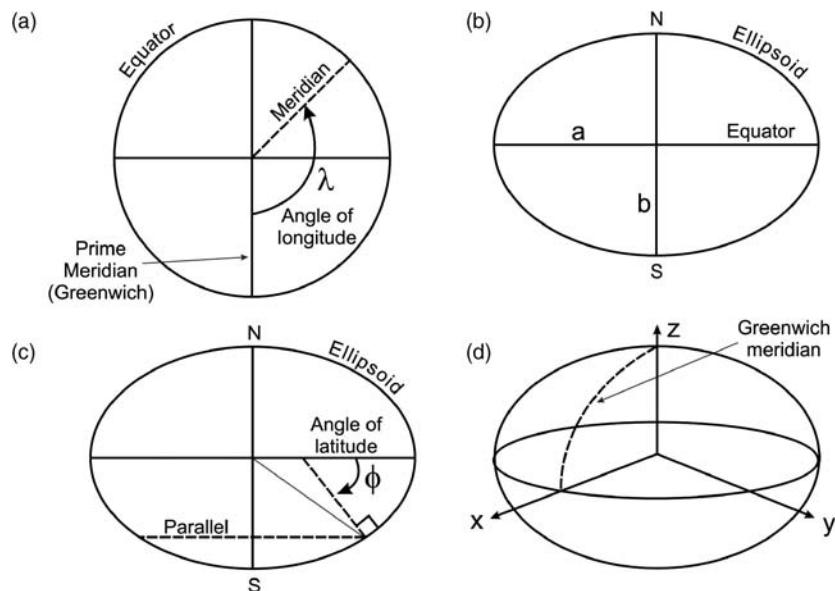
To define the centre of mass and so latitude and longitude (see Figure 13.1), we must first define the Earth's *axis of rotation* and the plane through the centre of mass perpendicular to the axis (the *equator*). Slices parallel to the axis but perpendicular to the plane of the equator are lines of constant longitude; these pass through the centre of mass and are sometimes also referred to as *great circles*. The slice through Greenwich defines zero degrees

longitude and the angle between it and any other slice defines the angle of longitude, so that longitude then goes from  $180^\circ$  west to  $180^\circ$  east of Greenwich. A line of constant longitude is also called a *meridian*. Perpendicular to a meridian, a slice perpendicular to the axis and passing through the Earth but not through its centre is called a *parallel*, also referred to as a small circle, except for the equator (which is a great circle).

We also need to describe the shape of the Earth, and the best approximation of this is the *ellipsoid of rotation* or *spheroid*. An ellipsoid is a type of quadric surface and is the 3D equivalent of an ellipse. It is defined, using  $x$ ,  $y$ ,  $z$  Cartesian coordinates, by

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1. \quad (13.1)$$

The Earth is not spherical but oblate and so the difference between the ellipsoid or spheroid and a perfect sphere is defined by its *flattening* ( $f$ ), or its reduction in the shorter *minor axis* relative to the *major axis*. *Eccentricity* ( $e$ ) is a further phenomenon which describes how the shape of an ellipsoid



**Figure 13.1** Schematic representation of the Earth, looking (a) down the pole, perpendicular to the equator, (b) and (c) perpendicular to the pole from the equator, and (d) obliquely at the Earth. This illustrates the relationship between longitude and the meridians (a), between the equator and major and minor semi-axes (b), and between latitude and parallels (c), and the locations of the  $x$ ,  $y$  and  $z$  axes forming Cartesian coordinates (d)

deviates from a sphere (the eccentricity of a circle being zero). Flattening and eccentricity then have the following relationships:

$$f = \frac{(a - b)}{a} \text{ and } e^2 = \frac{a^2 - b^2}{a^2} \text{ or } e^2 = 2f - f^2 \quad (13.2)$$

where  $a$  and  $b$  are the lengths of the major and minor axes respectively (usually referred to as semi-axes or half lengths of the axes). The actual flattening for the Earth's case is about 1 part in 300. Some of the first ellipsoids to be established were not particularly accurate and were not actually centred on the Earth's centre of mass. Fortunately, and rather ironically, the Cold War, the nuclear arms race and the need to target intercontinental missiles helped to drive the development of an international standard ellipsoid. The ellipsoid known as the World Geodetic System of 1984 (or WGS84) is now accepted as this standard although many others are in use.

*Latitude* can now be defined as the angle between the equator and a line perpendicular to the ellipsoid, which ranges from  $90^\circ$  north or south of the equator (see Figure 13.1c). Latitude is commonly given the Greek symbol phi ( $\phi$ ) and longitude lambda ( $\lambda$ ). A line of constant latitude is known as a parallel. Parallels never meet since they are parallel to one another, whereas meridians (lines of longitude) converge at the poles.

*Longitude* is more complex and only east–west measurements made at the equator are true. Away from the equator, where the lines of latitude decrease in length, measures are increasingly shortened, by approximately the cosine of the latitude. This means that at  $30^\circ$  north (or south), shortening is about 0.866, 0.707 at  $45^\circ$  and 0.5 at  $60^\circ$ . At  $60^\circ$  north

or south,  $1^\circ$  of longitude will represent 55 km ground distance.

### 13.2.2 Measuring height: the geoid

The true shape of the Earth forms a surface which is perpendicular to the direction of gravity, and is described as an *equipotential surface*, in which there are fluctuations and irregularities according to variations in the density of the crust and mantle beneath. The spheroid or ellipsoid can therefore be thought of as a reasonable representation of the shape of the Earth but not the true shape; this we refer to as the *geoid*, and it is defined as an 'equipotential surface which most closely resembles mean sea level'. Mean sea level is used in this context since it refers to everywhere and is a surface perpendicular to gravity. In general, differences between mean sea level and the geoid (referred to as *separation*) are greatest where undulations in the terrain surface are of the greatest magnitude but are generally less than 1 m.

The significance of the geoid's variability is that it leads to different definitions of height from one place to another, since mean sea level also varies. Different countries may define slightly different equipotential surfaces as their reference. We should therefore take some care to distinguish between heights above geoid or spheroid (Figure 13.2). Fortunately the differences are small so that only highly precise engineering applications should be affected by them. For reference, *orthometric heights* and *spheroidal heights* are those defined with respect to the geoid and spheroid respectively. The variation in height from the geoid gives us topography.

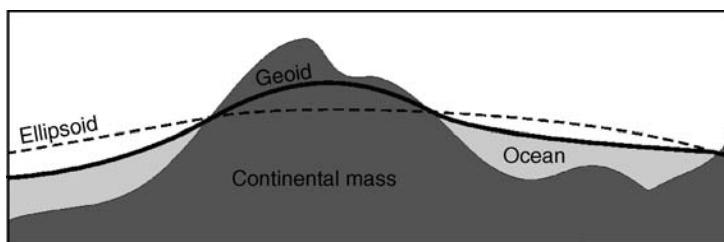


Figure 13.2 Heights above the spheroid and geoid

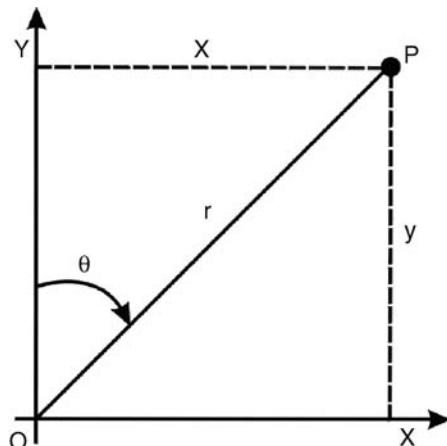
### 13.2.3 Coordinate systems

Many different generic types of coordinate systems can be defined and the calculations necessary to move between them may sometimes be rather complex. In order of increasing complexity they can be thought of as follows. *Spherical coordinates* are formed using the simplest approximation of a spherical Earth, where latitude is the angle north or south of the equatorial plane, longitude is the angle east or west of the prime meridian (Greenwich) and height is measured above or below the surface of the sphere. If high accuracy is not important, then this simple model may be sufficient for your purposes. *Spheroidal coordinates* are formed using a better approximation based on an ellipsoid or spheroid, as described in 13.2.1, with coordinates of true latitude, longitude and height; this gives us the system of *geodetic coordinates*. *Cartesian coordinates* involve values of  $x$ ,  $y$  and  $z$ , and are defined with their origin at the centre of a spheroid (see Figure 13.1d). The  $x$  and  $y$  axes lie in the equatorial plane, with  $x$  aligned with the Greenwich meridian, and  $z$  aligned with the polar axis. *Projection coordinates* are then defined using a simple set of  $x$  and  $y$  axes, where the curved surface of the Earth is transformed onto a plane, the process of which causes distortions (discussed later).

*Polar coordinates*, generically, are those which are defined by distance and angle, with distance usually denoted  $r$  and angle  $\theta$ . *Planar coordinates* refer to the representation of positions, as identified from polar coordinate positions, on a plane within which a set of orthogonal  $x$ ,  $y$  axes is defined. The conversion between these polar and planar coordinates, for any particular datum, is relatively straightforward and the relationship between them is illustrated in Figure 13.3. The following expression can be used to derive the distance ( $d$ ) between two points  $a$  and  $b$  on an assumed spherical Earth:

$$d(a, b) = R \arcsin[\sin \phi_A \sin \phi_B + \cos \phi_A \cos \phi_B \cos(\lambda_B - \lambda_A)] \quad (13.3)$$

where  $R$  is the radius of the Earth,  $A$  and  $B$  denote the positions of points  $a$  and  $b$  on the sphere,  $\lambda$  is the longitude and  $\phi$  the latitude. Generically, the  $x$ ,  $y$  (planar) positions of the two points can be



**Figure 13.3** The relationship between polar and planar coordinates in a single datum

derived from the polar coordinates as

$$x = r \sin \theta \text{ and } y = r \cos \theta \quad (13.4)$$

$$r = \sqrt{x^2 + y^2} \text{ and } \theta = \arctan\left(\frac{y}{x}\right) \quad (13.5)$$

where  $\theta$  is measured clockwise from north. The Pythagorean distance between two points ( $a$  and  $b$ ) can then be found by the following, where the two points are located at  $(x_a, y_a)$  and  $(x_b, y_b)$ :

$$d(a, b) = \sqrt{(x_a - x_b)^2 + (y_a - y_b)^2}. \quad (13.6)$$

### 13.2.4 Datums

A *geodetic datum* is a mathematical approximation of the Earth's 3D surface and a reference from which other measurements are made. Every spheroid has a major axis and a minor axis, with the major axis being the longer of the two (as shown in Figure 13.1) but is not in itself a datum. The missing information is a description of how and where the shape deviates from the Earth's actual surface. This is provided by the definition of a *tie point*, which is a known position on the Earth's surface (or its interior, since the Earth's centre of mass could be used), and its corresponding location on or within the ellipsoid.

Complications arise because datums may be global, regional or local, so that each is only accurate for a limited set of conditions. For a *global datum*, the tie point may well be the centre of mass of the Earth, meaning that the ellipsoid forms the best general approximation of the Earth's shape, and that at any specific positions and accuracies may be quite poor. Such generalizations would be acceptable for datasets which are of very large or global extent. In contrast, a *local datum*, which uses a specific tie point somewhere on the surface, near the area of interest, would be used for a 'local' projector or dataset. Within this area, the deviation of the ellipsoid from the actual surface will be minimal but at some distance from it may be considerable. This is the reason behind the development of the great number of datums and projections worldwide. In practice, we choose a datum which is appropriate to our needs, according to the size and location of the area we are working with, to provide us with optimum measurement accuracy. Some common examples are given in Table 13.1.

It is worth noting that in many cases there may be several different versions of datum and ellipsoid under the same name, depending on when, where, by whom and for which purpose they were developed. The differences between them may seem insignificant at first glance but, in terms of calculated ground distances, could produce very significant differences between measurements.

### 13.2.5 Geometric distortions and projection models

Since paper maps and geospatial databases are flat representations of data located on a curved surface,

the *map projection* is an accepted means for fitting all or part of that curved surface to the flat surface or plane. This projection cannot be made without distortion of shape, area, distance, direction or scale. We would ideally like to preserve all these characteristics but we cannot, so we must choose which of them should be represented accurately at the expense of others, or whether to compromise on several characteristics.

There are probably 20 or 30 different types of map projections in common usage. These have been constructed to preserve one or other characteristics of geometry, as follows:

1. *Area*: Many map projections try to preserve area, so that the projected region covers exactly the same area of the Earth's surface no matter where it is placed on the map. To achieve this the map must distort scale, angles and shape.
2. *Shape*: There are two groups of projections which have either:
  - (a) a conformal property where the angles and the shapes of small features are preserved, and the scales in  $x$  and  $y$  are always equal (although large shapes will be distorted); or
  - (b) an equal area property where the areas measured on the map are always in the same proportion to the areas measured on the Earth's surface but their shapes may be distorted.
3. *Scale*: No map projection shows scale correctly everywhere on the map, but for many projections there are one or more lines on the map where scale is correct.

**Table 13.1** Examples of geodetic datums and their descriptive parameters

Datum	Spheroid	$a$	$b$	$f$	Tie point
National Grid of Great Britain (OSGB)	Airy 1830	6 377 563	6 356 256.9	1/299.32	Herstmonceux
Pulkovo <sup>a</sup>	Krassovsky 1940	6 378 245	6 356 863	1/298.3	Pulkovo Observatory
M'poraloko 1951 <sup>a</sup>	Clarke 1880	6 378 249.2	6 356 515.0	1/293.47	Libreville

<sup>a</sup>Indicates that more than one variant exists for this datum.

4. *Distance*: Some projections preserve neither angular nor area relationships but distances in certain directions are preserved.
5. *Angle*: Although conformal projections preserve local angles, one class of projections (called *azimuthal projections*) preserve the easting and northing pair, so that angle and direction are preserved.

Scale factor ( $k$ ) is useful to quantify the amount of distortion caused by projection and is defined by the following:

$$k = \frac{\text{Projected distance}}{\text{Distance on the sphere}}. \quad (13.7)$$

This relationship will be different at every point on the map and in many cases will be different in each direction. It only applies to short distances. The ideal scale factor is 1, which represents no distortion at all. Most scale factors approach but are less than 1.

Any projection can achieve one or other of these properties but none can preserve all, and the distortions that occur in each case are illustrated schematically in Figure 13.4.

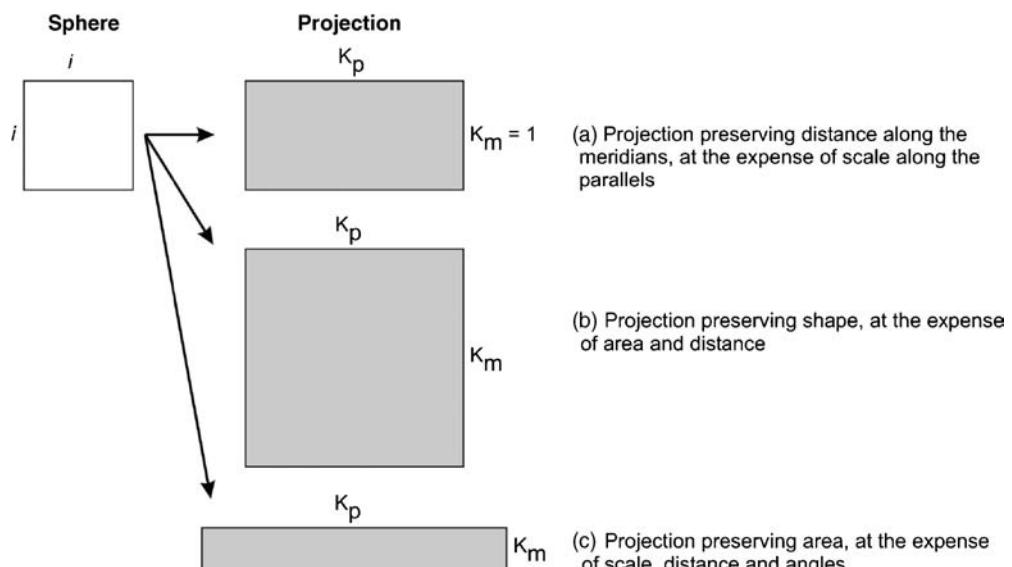
Once the property to be preserved has been decided, the next step is to transform the information

using a ‘projectable’ or ‘flattenable’ surface. The transformation or projection is achieved using planar, cylindrical or cone-shaped surfaces that touch the Earth in one of a few ways (Figure 13.5); these form the basis for the three main groups of projection. Where the surface touches the Earth at a point (for a plane), along a great circle (for a cylinder) or at a parallel (for a cone), projections of the *tangent* type are formed. Where the surface *cuts* the earth, rather than just touching it, between two parallels, a *secant* type of projection is formed (see Figure 13.6). The conic is actually a general form, with azimuthal and cylindrical forms being special cases of the conic type.

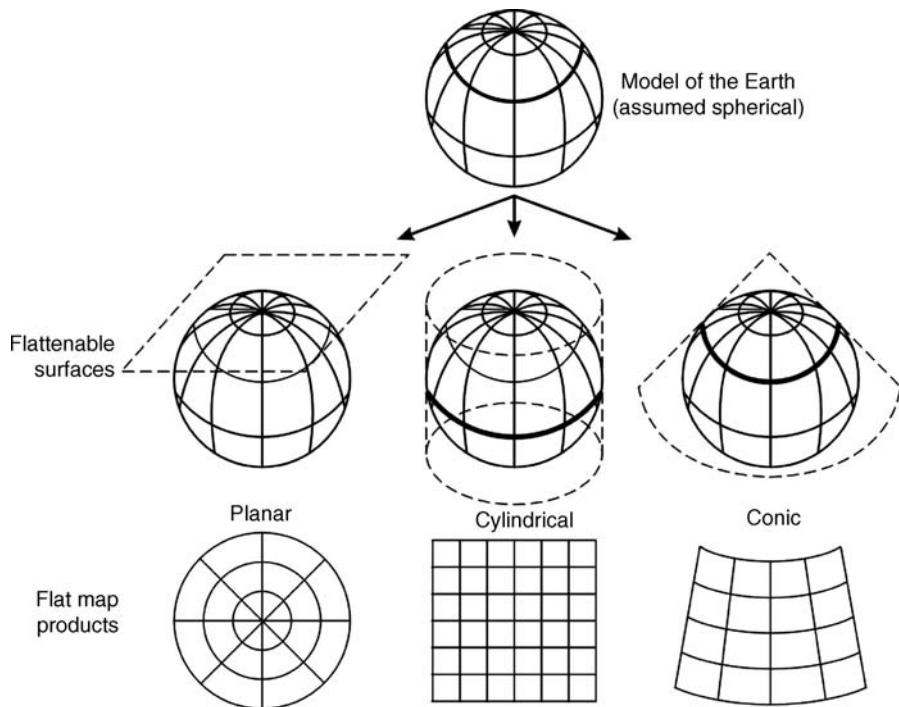
In the planar type (such as a stereographic projection) where the surface is taken to be the tangent to one of the poles, the following relationship can be used to derive polar positions (all the equations given here assume a spherical Earth for simplicity):

$$\theta = \lambda \text{ and } r = 2 \tan\left(\frac{\chi}{2}\right) \quad (13.8)$$

where  $\chi$  represents the colatitude ( $\chi = 90^\circ - \phi$ ); the resultant polar coordinates can then be converted to planar coordinates using Equation (13.3). Of course, the plane could be a tangent to the Earth at any point, not just at one of the poles.



**Figure 13.4** Schematic illustration of the projection effects on a unit square of side length  $i$ , where  $K$  represents the scale along each projected side, and subscripts m and p represent meridian and parallel



**Figure 13.5** The three main types of projection which are based on the tangent case: planar (left), cylindrical (centre) and conic (right). Modified after Bonham-Carter (2002)

For cylindrical projections, the axis of the cylinder may pass through the poles, for example, so that it touches the sphere at the equator (as in the Mercator). In this case, positions may be derived as

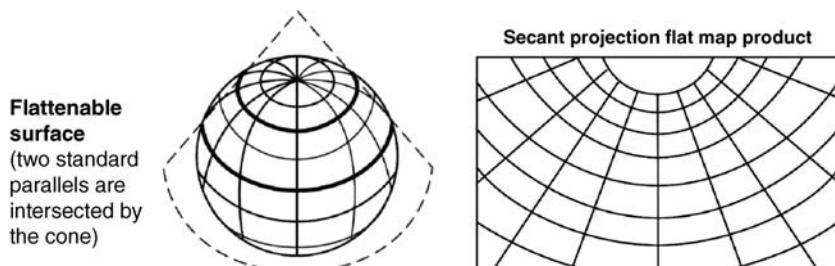
$$x = \lambda \text{ and } y = \log_e \tan\left(\frac{\pi}{4} + \frac{\phi}{2}\right). \quad (13.9)$$

For conic projections of the tangent type, the following can be used to derive positions, assuming one standard parallel at a colatitude of  $\chi_0$ :

$$r = \theta = \lambda \cos(\chi_0) \text{ and } \tan(\chi_0) + \tan(\chi - \chi_0). \quad (13.10)$$

### 13.2.6 Major map projections

Based on the projection models described so far, there are three broad categories of projection: equidistant, equal area and conformal.



**Figure 13.6** The conic projection when based on the secant case, where the conic surface intersects at two standard parallels. Modified after Bonham-Carter (2002)

### 13.2.6.1 Cylindrical equidistant projection

In this projection the distances between one or two points and all other points on the map differ from the corresponding distances on the sphere by a constant scaling factor. The meridians are straight and parallel, and distances along them are undistorted. Scale along the equator is true but on other parallels is increasingly distorted towards the poles; shape, area are therefore increasingly distorted in this direction. An example is shown in Figure 13.7a. The *Plate Carrée* is an example of this type of projection.

### 13.2.6.2 Cylindrical equal area projection

Here scale factor is a function of latitude so that, away from the equator, distances cannot be measured from a map with this type of projection. Shape distortion is extreme towards the poles. Scale factor is almost 1 on both meridians and parallels, as follows, but only near the equator:

$$k_m k_p = 1. \quad (13.11)$$

Scale factor along the parallels is given by  $\sec \phi$  and distortion along the meridians by  $\cos \phi$ . An example is the *Lambert cylindrical equal area projection* (e.g. Figure 13.7b).

### 13.2.6.3 Conformal: Mercator

In this case, the poles reach infinite size and distance from the equator, producing a very distinct-looking map. The meridians are parallel and angles are preserved so these maps are acceptable for navigation (illustrated in Figures 13.7c and 13.8). The scale factor ( $k$ ) at any point on the map is a function of latitude, as follows:

$$k_p = k_m = \sec \phi \quad (13.12)$$

The distortion that occurs towards (and which is infinite at) the poles is a product of the way the Mercator is constructed, as illustrated in Figure 13.8a. This can be thought of by considering an area between two meridians, both at the equator and at a high latitude. The equatorial area appears square with sides of  $1^\circ$  length in both longitude and latitude. At high latitudes the area covers the same distance in latitude as at the equator and is still  $1^\circ$  wide but is narrower and covers a distance shown by  $x$  in Figure 13.8, so the area is no longer square but

rectangular. On the projected map (Figure 13.8b) the two meridians are shown, representing a difference of  $1^\circ$  longitude, and they are parallel, so when the rectangle is projected it must be enlarged by a factor  $1/x$  to fit between them. This is the reason why areas are larger at higher latitudes and when this scaling is done repeatedly, from the equator northwards, the Mercator coordinate net is produced. On the Earth, a  $1^\circ$  distance at the equator is about 111 km.

### 13.2.6.4 Conformal: transverse mercator

This projection is a modification of the standard Mercator designed for areas away from the equator. In this instance, the cylinder is rotated through  $90^\circ$  so that its axis is horizontal and its point of contact with the Earth is no longer the equator but a meridian. The scale factor is same in any direction, and is defined by

$$k = \sec \theta \quad (13.13)$$

where  $\theta$  is equivalent to latitude ( $\phi$ ), except that it is the angular distance from the central meridian rather than from the equator. In this case the meridians are no longer parallel or straight (except the central meridian), and the angle made between the meridians and the central meridian (which is grid north) can be described as *convergence* ( $\gamma$ ). For the sphere projection convergence is defined by

$$\gamma = \delta \lambda \sin \phi. \quad (13.14)$$

Then by turning the cylinder horizontally, the central meridian of a transverse Mercator (TM) projection could be based on any line of longitude around the globe so as to be appropriate to any particular country or region. The British National Grid (OSGB) is a good example of this type of projection: the United Kingdom covers a greater distance in the north-south direction than it does in the east-west direction, so to prevent north-south distortion (as would be produced by a Mercator), a TM with its central meridian at  $0^\circ$  longitude (i.e. Greenwich) is used (see Figure 13.7d). Further rules can then be applied to the TM to produce equal area, equidistant or conformal projections.

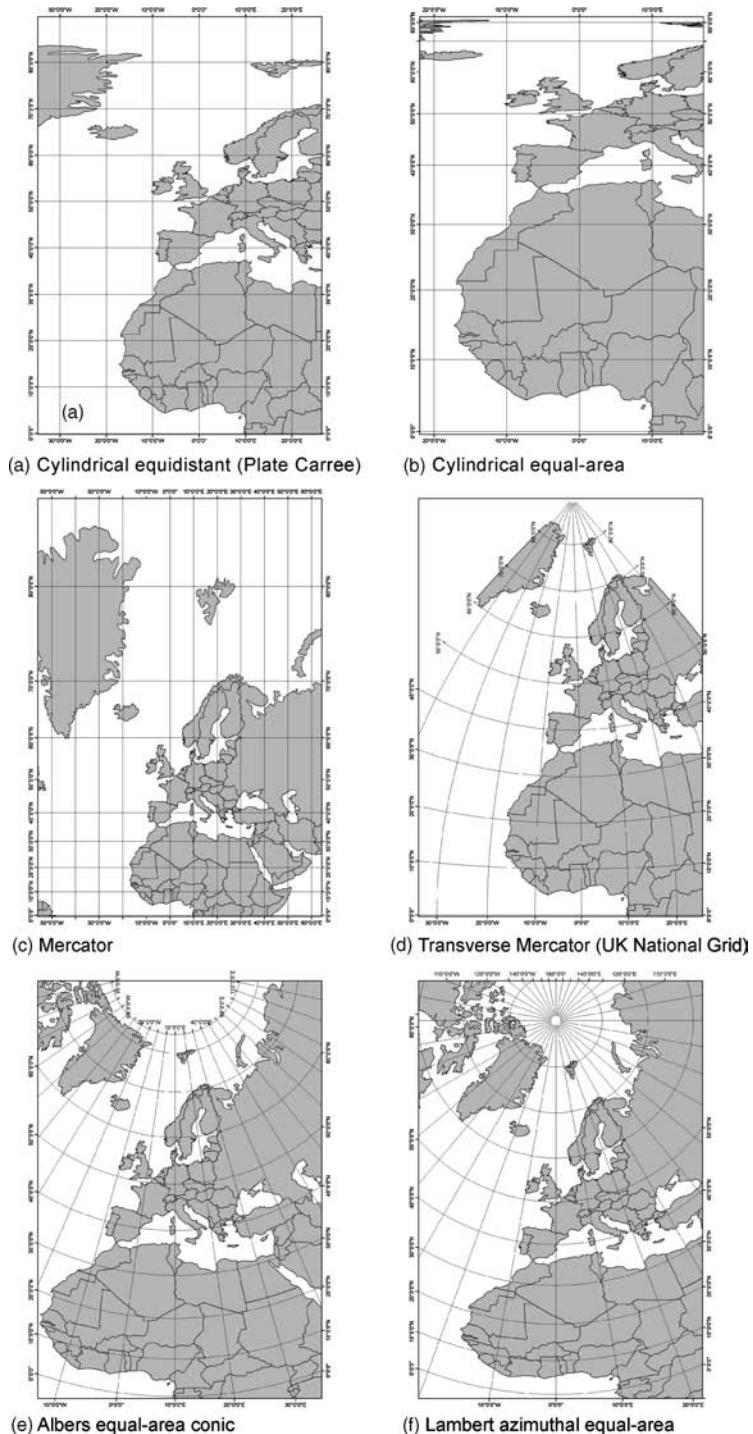
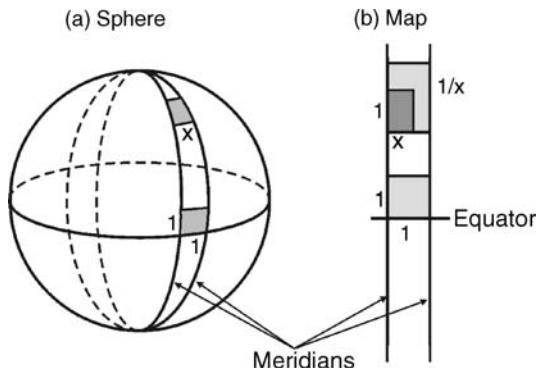


Figure 13.7 Common projection examples, as described in the text



**Figure 13.8** The relationship between meridians and scale on the sphere (a) and in the Mercator projection (b)

#### 13.2.6.5 Conic projections

In these projections, the meridians are straight, have equal length and converge on a point that may or may not be a pole. The parallels are complete concentric circles about the centre of the projection. Such projections may be conformal, equidistant or equal area. Examples include the *Albers equal area conic projection* (illustrated in Figure 13.7e).

#### 13.2.6.6 Planar (azimuthal) projections

Azimuthal projections preserve the azimuth or direction from a reference point (the centre of the map) to all other points; i.e., angles are preserved at the expense of area and shape. In the polar case, the meridians are straight and the parallels are complete concentric circles. Scale is true only near the centre, with the map being circular in form. Such projections may be conformal, equidistant or equal area. Examples include the *Lambert Azimuthal Equal Area projection* (Figure 13.7f), or the *stereographic azimuthal projection*.

#### 13.2.6.7 Conformal: Universal Transverse Mercator

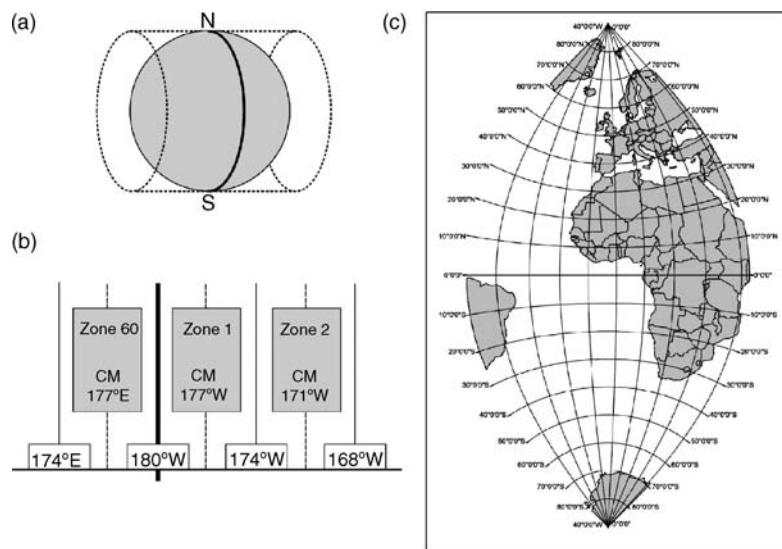
A further modification of the Mercator allows the production of the Universal Transverse Mercator (UTM) projection system (illustrated in Figure 13.9). It is again projected on a cylinder tangent to a meridian (as in the TM) and by repeatedly turning the cylinder, about its polar axis, the world can be divided into 60 east–west zones, each  $6^\circ$  longitude in width. The projection is conformal so that shapes and angles within any small area will be preserved. This system was originally adopted for large-scale

military maps for the world, but it is now a global standard and again is useful for mapping large areas that are oriented in a north–south direction.

Projected UTM grid coordinates are then established which are identical between zones. Separate grids are also established for both northern and southern halves of each UTM zone to ensure there are no negative northings in the southern hemisphere. Hence, when quoting a UTM grid reference, it is essential to state eastings, northings, zone number and the hemisphere (north or south) to ensure clarity.

#### 13.2.7 Projection specification

Several other parameters are necessary to define precisely a particular projection. For example, the Mercator requires a central meridian to be specified, and this is given as the line of longitude on which the projection is centred (it is commonly expressed in radians). Coordinates defining a *false northing and easting* are also required, to position the origin of the projection. These are normally given in metres and are used to ensure that negative values are never encountered. A *scale factor* is also given, as a number by which all easting and northing values must be multiplied, to force the map onto the page. A conic projection of the secant type requires the specification of the *first and second parallel* (as latitudes thereof). Some projections also require the origin of the projection to be given as a latitude and longitude coordinate pair; these are sometimes referred to as the *central meridian* and *central parallel* (Table 13.2).



**Figure 13.9** The Universal Transverse Mercator system: (a) the cylindrical projection rotated about its axis, touching the Earth along a meridian (any meridian) rather than the equator; (b) the arrangement of  $6^{\circ}$  zones and their central meridians, starting at  $180^{\circ}\text{W}$  (there are 60 zones in all); (c) the arrangement of parallels and meridians, in UTM zone 30, about the central meridian, which in this case is  $0^{\circ}$  at Greenwich, appropriate for the United Kingdom, western Europe and north-west Africa (separate projected grid coordinates are established for the northern and southern halves of each UTM zone so that negative northings are never encountered)

The *geodetic* projection is a special type of map projection (the simplest possible map projection) where the easting value is exactly longitude, and the northing value is exactly latitude. Since it does not preserve angle, scale or shape, it is not generally used for cartographic purposes but is commonly used as the default option for recording simple coordinate positions since it is the only globally constant system. The datum associated with this projection, to preserve its global applicability, is always the *World Geodetic System 1984* (WGS84).

### 13.3 How coordinate information is stored and accessed

Vector data store their coordinate information implicitly with each node position in real-world coordinates and these are used directly to plot positions and to re-project from one coordinate system to another. Raster data on the other hand have a regular local row and column number system for each pixel so that, internally, the geometry and position of a

**Table 13.2** Some examples of projected coordinate systems

Projection	Central meridian (longitude)	Central parallel (latitude)	False easting	False northing	Scale factor
OS GB	$2^{\circ}\text{W}$	$49^{\circ}\text{N}$	+ 400 000	- 100 000	0.999 601 2
UTM $\phi > 0^{\circ}$	Zonal	$0^{\circ}$	+ 500 000	0	0.999 6
UTM $\phi < 0^{\circ}$	Zonal	$0^{\circ}$	+ 500 000	+ 100 000	0.999 6
GK TM zone 11	$63^{\circ}\text{E}$	$0^{\circ}$	+ 500 000	0	1

pixel's location are implicit. Externally, however, a world geographic reference must be explicitly stated; this requires the geographic location of the image origin and the ground distance represented by each pixel. A transformation is then performed, which converts local image coordinates to real-world coordinates for each pixel location using the geometric operations described in Chapter 9. This transformation information is also stored explicitly.

There are many binary image formats, such as IMG, BSQ, BIL, BIP, GeoTIFF and various ASCII grid formats, which are now accepted as standards. Some of these store the georeferencing information in the header portion of the actual image data file. Other image formats store the information in a separate ASCII text file, sometimes referred to as the *header file* or *world file*, since it contains the real-world transformation information used by the image. Since these files are in ASCII text format, they can be created or edited with any text editor. Most GIS/mapping/CAD software will detect and read this information automatically, if it is present. The image-to-world transformation is accessed each time an image is displayed and visualized.

The contents of a world file, for a projected raster image with plane coordinates, will look something like this:

10.000	(A)
0.00000000000000	(B)
0.00000000000000	(C)
-10.000	(D)
567110.113454530548	(E)
9415540.445499603346	(F)

Most GIS software, when this file is present, will perform an affine transformation of the following form:

$$x_1 = Ax + Cy + E \text{ and } y_1 = Bx + Dy + F \quad (13.15)$$

where  $x_1$  and  $y_1$  are the calculated coordinates of a pixel on the map,  $x$  and  $y$  are the pixel column and row numbers in the image,  $A$  is the  $x$  scale (dimension of a pixel in map units in the  $x$  direction),

$D$  is the negative of the  $y$  scale (dimension of a pixel in map units in the  $y$  direction),  $B$  and  $C$  are rotational terms, and  $E$  and  $F$  are translational terms (centre coordinates of the top-left pixel). Note that the  $y$  scale ( $D$ ) is negative. This is because the origins of a geographic coordinate system and of a raster image are different: the geographic origin is usually in the lower left corner whereas the origin of an image is the upper left corner, so that  $y$  coordinate values in the map increase from the origin upwards and raster row values increase from the origin downwards.

Programs such as ER Mapper, ArcInfo (ArcGIS), ENVI, ERDAS Imagine and PCI Geomatica contain routines to convert both image and vector data between projections and datums, while some programs only support re-projection of vector data.

## 13.4 Selecting appropriate coordinate systems

It is common to receive geospatial data created or acquired by someone else, and we frequently need to overlay datasets that are in different or unknown or unspecified coordinate systems. Establishing the projection and datum becomes of vital importance before beginning any work and here again the use and upkeep of metadata is vital. The metadata should always contain information on the datum and projection of a geospatial dataset in addition to other information necessary to document its provenance. Such information is often found in the header file (or metadata) in the case of raster images, or failing that, the dataset may have to be visually compared with another dataset of known datum and projection.

When creating a new dataset, or defining a new project, selection of the most appropriate map projection for any input data it is very important. Selection considerations include the relative size of the project area (e.g. the world, a continent or a small region), its location (e.g. polar, mid-latitude or equatorial) and its predominant extent (e.g. circular, east–west axis, north–south axis or oblique axis). If there is a pre-existing base layer, such as a scaled map or georectified image, this may form the

framework for all other data that are added or created within the project.

The UTM projection is nearly correct in every respect, for relatively small project areas, and is a very common choice. There are some general ‘rules of thumb’ which are useful for continental-scale, or smaller, regions in mid-latitude zones. For instance, for areas with a dominantly north–south axis, UTM will provide conformal accuracy; with an east–west axis, selection of a Lambert conformal projection will give conformal accuracy or the Albers equal area projection will preserve area. Areas with an oblique axis could be represented well by an oblique Mercator projection (for conformal accuracy), and for an area that has equal extent in all directions, a polar or stereographic projection will give conformal accuracy, or a Lambert azimuthal projection could be chosen for area preservation.

The standard Gauss–Kruger (GK) projection is sometimes also known as the Pulkovo 1942 Gauss–Kruger projection. Gauss–Kruger projections are implemented as a National Grid in Germany, referred to as the DHDN Gauss–Kruger, and are also commonly used in Russia and China. The GK projection is particularly suited for this part of the world since these countries occupy large continental masses of considerable east–west

extent. A zonal system, similar to that of the UTM but with zones of  $3^{\circ}$  width instead of  $6^{\circ}$ , is used for Russia and China, to ensure minimal distortion and maximum conformality across the continent.

## Questions

- 13.1 Why is it important to establish a structural framework for the representation of digital data?
- 13.2 What properties of the spherical Earth are affected by the use of map projections?
- 13.3 How would you decide on the most appropriate framework for a project?
- 13.4 What are the advantages and disadvantages of the UTM system of reference?
- 13.5 What happens to the area represented by a pixel when it is transformed between geodetic and projected coordinates, particularly at high latitudes?
- 13.6 How have you recorded your geographic location data (field localities) in the past? How accurately did you record their positions? What coordinate system did you choose? And how would you choose to record and display them now?



# 14

## Operations

### 14.1 Introducing operations on spatial data

It is probably fair to say that the average GIS suite contains far more functions than most people will ever need or be aware of! A long list of these, with some descriptions, would be useful but would make rather dull reading, so it is helpful for the purposes of understanding to categorize them in some way.

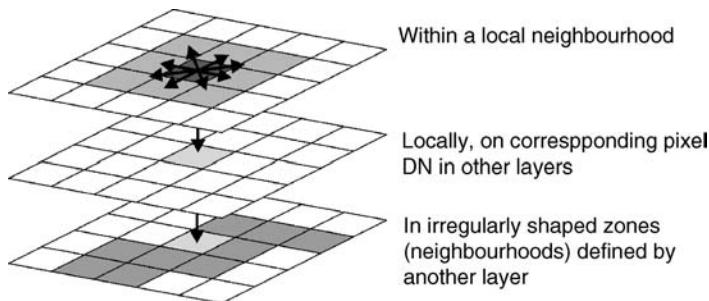
The way that processes are carried out depends on how the data are structured and stored and since we have already described the fundamental differences between vector and raster data (in Chapter 12) we should begin to understand the implications of using one method of description over another. The difference between operations on raster and vector data could be thought of as a dimensional one. Since one raster represents the variance of one attribute, operations affecting one raster pixel value occur in a ‘vertical’ dimension, through the stack of raster data attributes, or a ‘horizontal’ dimension on one attribute only (Figure 14.1), or a combination of these two. Operations on vector features’ attribute values occur in an  $n$ -dimensional space since the features’ values are stored in an attribute table which has  $n$  attribute fields and the spatial extents of the input features are neither regular nor necessarily equal.

Operations could be grouped on the basis of being either ‘spatial’ or ‘non-spatial’. Those falling into the non-spatial category could include reclassification

operations or statistical operations carried out within tables. Truly spatial operations could include neighbourhood processes such as convolution filtering (see also Chapter 4), or functions used to enhance the contrast of a raster image since these involve the statistics of a region or of the whole image (see also Chapter 3). In these cases, the processing itself involves the manipulation of data in a spatial context and produces results that reveal spatial patterns more clearly.

Since the objectives of many ‘non-spatial’ operations may also lead to and be part of wider spatial analyses, it is perhaps more useful to describe them as operations which are carried out on spatial data. Clearly there is a grey area here, and this is the reason for referring simply to operations that are performed upon data which are spatial in nature rather than classifying the operations themselves.

A further, rather useful hybrid classification of analytical operations could be made on the basis of the type of output, map or table, as well as on whether spatial variables were involved or not, as summarized in Table 14.1, such that the simple reassignment of values in one raster to another scheme of values in a new raster (i.e. reclassification) could be considered to produce map output but not necessarily involve spatial attributes; or, at the opposite end of the spectrum, the calculation of spatial autocorrelation to produce a variogram provides tabular output and definitely involves the use of attributes with spatial qualities.



**Figure 14.1** Stacked georeferenced rasters indicating multi-layer operations, both local and neighbourhood

Another, perhaps more instructive way of classifying them is in terms of the number of input data ‘layers’ involved: one, two or more (Figure 14.1). In general, those operations applied to raster data are essentially the same as image processing, and here is the overlap between image processing and GIS. For instance, multi-layer raster data operations are no different from multi-spectral image operations except the attributes carry different meaning.

brought to GIS, map algebra provides an elegant way to describe operations on GIS datasets. It can be thought of simply as algebra applied to spatial data which, in the case of raster data, are facilitated by the fact that a raster is a georeferenced numerical array.

## 14.2 Map algebra concepts

*Map algebra* is an informal and commonly used scheme for manipulating continuously sampled (i.e. raster) variables defined over a common area. It is also a term used to describe calculations within and between GIS data layers, according to some mathematical expression, to produce a new layer; it was first described and developed by Tomlin (1990). Map algebra can also be used to manipulate vector map layers, sometimes resulting in the production of a raster output. Although no new capabilities are

### 14.2.1 Working with null data

An essential part of map algebra or spatial analysis is the coding of data in such a way as to eliminate certain areas from further contribution to the analysis. For instance, if the existence of low-grade land is a prerequisite for a site selection procedure, we then need to produce a layer in which areas of low-grade land are coded distinctively so that all other areas can be removed. One possibility is to set the areas of low-grade land to a value of 1 and the remaining areas to 0. Any processes involving multiplication, division or geometric mean that encounter the zero value will then also return a zero value and that location (pixel) will be removed from the

**Table 14.1** Operations categorized according to their spatial or non-spatial nature. After Bonham-Carter (2002)

Spatial attributes involved?		
Output	Yes	No (not necessarily)
Map or image	Neighbourhood processing (filtering), zonal and focal operations, mathematical morphology	Reclassification, rescaling (unary operations), overlay (binary operations), thresholding and density slicing
Tabular	Spatial autocorrelation and variograms	Various tabular statistics (aggregation, variety) and tabular modelling (calculation of new fields from existing ones), scattergraphs

analysis. The opposite is true if processing involves addition, subtraction or arithmetic mean calculations, since the zero value will survive through to the end of the process. The second possibility is to use a null or NoData value instead of a zero. The null is a special value which indicates that no information is associated with the pixel position, i.e. there is no digital numerical value. In general, unlike zero, any expression will produce a null value if any of the corresponding input pixels have null values.

Many functions and expressions simply ignore null values, however, and in some circumstances this may be useful, but it also means that a special kind of function must be used if we need to test for the presence of (or to assign) null values in a dataset. For instance, within ESRI's ArcGIS, the function *ISNULL* is used to test for the existence of null values and will produce a value of 1 if null, or 0 if not. Using ER Mapper's formula editor, null values can easily be assigned, set to other values, made visible or hidden. Situations where the presence of nulls is disadvantageous include instances where there are unknown gaps in the dataset, perhaps produced by measurement error or failure. Within map algebra, however, the null value can be used to great advantage since it enables the selective removal or retention of values and locations during analysis.

### 14.2.2 Logical and conditional processing

These two processes are quite similar and they provide a means of controlling what happens during some function. They allow us to evaluate some criterion and to specify what happens next if the criterion is satisfied or not.

*Logical processing* describes the tracking of true and false values through a procedure. Normally, in map algebra, a non-zero value is always considered to be a logical true, and zero, a logical false. Some operators and functions may return either logical true values (1) or logical false values (0), for example relational and Boolean operators. The return of a true or false value acts as a switch for one or other consequence within the procedure.

*Conditional processing* allows that a particular action can be specified, according to the satisfaction of various conditions; if the conditions are

evaluated as true then one action is taken, and an alternative action is taken when the conditions are evaluated as false. The conventional *if-then-else* statement is a simple example of a conditional statement:

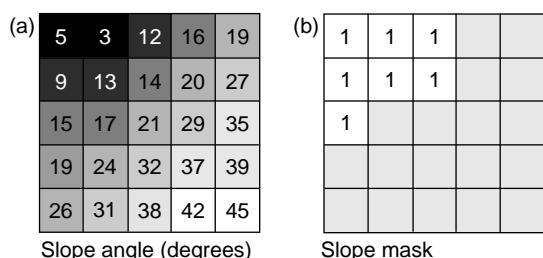
if  $i < 16$  then 1 else null where  $i = \text{input pixel } dn$

Conditional processing is especially useful for creating analysis 'masks'. In Figure 14.2, each input pixel value is tested for the condition of having a slope equal to or less than  $15^\circ$ . If the value tests true (slope angle is  $15^\circ$  or less), a value of 1 is assigned to the output pixel. If it tests false (exceeds  $15^\circ$ ), a null value is assigned to the output pixel. The output could then be used as a mask to exclude areas of steeper slopes and allow through all areas of gentle slopes, such as might be required in fulfilling the prescriptive criteria for a site selection exercise.

### 14.2.3 Other types of operator

Expressions can be evaluated using *arithmetic operators* (addition, subtraction, logarithmic, trigonometric) and performed on spatially coincident pixel DN values within two or more input layers (Table 14.2). Generally speaking, the order in which the input layers are listed denotes the precedence with which they are processed; the input or operator listed first is given top priority and is performed first, with decreasing priority from left to right.

A *relational operator* enables the construction of logical functions and tests by comparing two numbers and returning a true value (1) if the values are



**Figure 14.2** Logical test of slope angle data, for the condition of being no greater in value than  $15^\circ$ : (a) slope angle raster and (b) slope mask (pale grey blank cells indicate null values)

**Table 14.2** Summary of common arithmetic, relational, Boolean, power, logical and combinatorial operators

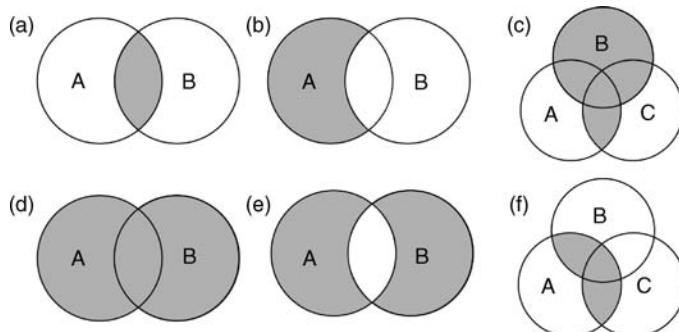
Arithmetic	Relational (return true/false)	Boolean (return true/false)
+, Addition −, Subtraction *, Multiplication /, Division MOD, Modulus	==, EQ Equal ^=, <>, NE Not equal <, LT Less than <=, LE Less than/equal to >, GT Greater than >=, GE Greater than/equal to	^, NOT Logical complement &, AND Logical AND  , OR Logical OR !, XOR Logical XOR
Power	Logical	Combinatorial
Sqrt, Square root Sqr, Square Pow, Raised to a power	DIFF, Logical difference IN {list}, Contained in list OVER, Replace	CAND, Combinatorial AND COR, Combinatorial OR CXOR, Combinatorial XOR

equal or false (0) if not. For example, this operator can be used to find locations within a single input layer with DN values representing a particular class of interest. These are particularly useful with discrete or categorical data.

A *Boolean operator*, for example AND, OR or NOT, also enables sequential logical functions and tests to be performed. Like relational operators, Boolean operators also return true (1) and false (0) values. They are performed on two or more input layers to select or remove values and locations from the analysis. For example, to satisfy criteria within a slope stability model, Boolean operators could be used to identify all locations where values in one input representing slope are greater than 40° AND where values in an elevation model layer are greater than 2000 m (as in Figure 14.3a).

*Logical operators* involve the logical comparison of the two inputs and assign a value according to the type of operator. For instance, for two inputs (A and B)  $A \text{ DIFF } B$  assigns the value from A to the output pixel if the values are different or a zero if they are the same. An expression  $A \text{ OVER } B$  assigns the value from A if a non-zero value exists; if not then the value from B is assigned to the output pixel. A *combinatorial operator* finds all the unique combinations of values among the attributes of multiple input rasters and assigns a unique value to each combination in the output layer. The output attribute will contain fields and attributes from all the input layers.

All these operators can be used, with care, alone or sequentially, to remove, test, process, retain or remove values (and locations) selectively from



**Figure 14.3** Use of Boolean rules and set theory within map algebra; here the circles represent the feature classes A, B and C, illustrating how simple Boolean rules can be applied to geographic datasets, and especially rasters to extract or retain values, to satisfy a series of criteria: (a) A AND B (intersection or minimum); (b) A NOT B; (c) (A AND C) OR B; (d) A OR B (union or maximum); (e) A XOR B; and (f) A AND (B OR C)

**Table 14.3** Summary of local operations

Type	Includes:	Examples
Primary	Creation of a layer from nothing	Rasters of constant value or containing randomly generated values
Unary	Conversion of units of measurement and as intermediary steps of spatial analysis	Rescaling, negation, comparing or applying mathematical functions, reclassification
Binary	Operations on ordered pairs of numbers in matching pixels between layers	Arithmetic and logical combinations of rasters
N-ary	Comparison of local statistics between several rasters (many to one or many to many)	Change or variety detection

datasets alone or from within a spatial analytical procedure.

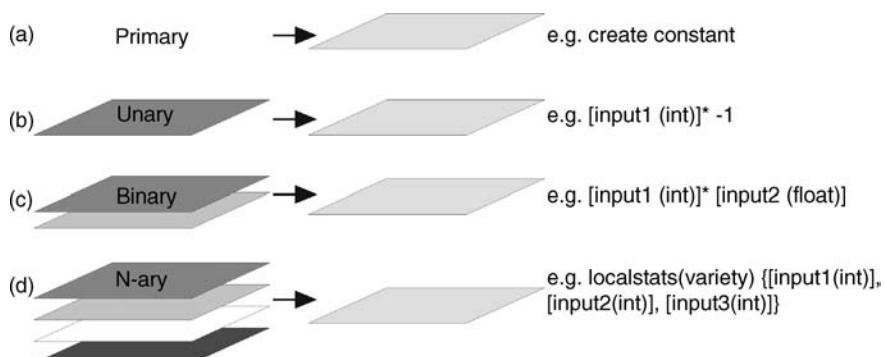
## 14.3 Local operations

A *local operation* involves the production of an output value as a function of the value(s) at the corresponding locations in the input layer(s). As described in Chapter 3, these operations can be considered point operations when performed on raster data, i.e. they operate on a pixel and its matching pixel position in other layers, as opposed to groups of neighbouring pixels, which are dealt with in Section 14.4. They can be grouped into those which derive statistics from multiple input layers (e.g. mean, median, minority), those which combine multiple input layers, those which identify values that satisfy specified criteria or the number of occurrences that satisfy specified criteria (e.g. greater than or less than), or those which identify the position in an input

list that satisfies a specified criterion. All types of operator previously mentioned can be used in this context. Commonly they are subdivided according to the number of input layers involved at the start of the process. They include primary operations where nothing exists at the start, to *n*-ary operations where *n* layers may be involved; they are summarized in Table 14.3 and illustrated in Figure 14.4.

### 14.3.1 Primary operations

This description refers primarily to operations used to generate a layer, conceptually from nothing, for example the creation of a raster of constant value, or containing randomly generated numbers, such as could be used to test for error propagation through some analysis. An output pixel size, extent, data type and output DN value (either constant or random between set limits) must be specified for the creation of such a new layer.

**Figure 14.4** Classifying map algebra operations in terms of the number of input layers and some examples

### 14.3.2 Unary operations

These operations act on one layer to produce a new output layer and they include tasks such as *rescaling*, *negation* and *reclassification*. Rescaling is especially useful in preparation for *multi-criteria analysis* where all the input layers should have consistent units and value range: for instance, in converting from byte data, with 0 to 255 value range, to a percentage scale (0–100) or a range of between 0 and 1, and vice versa. Negation is used in a similar context, in modifying the value range of a dataset from being entirely positive to entirely negative and vice versa. Reclassification is especially significant in data preparation for spatial analysis, and so deserves rather more in-depth description, but all these activities can be and are commonly carried out in image processing systems.

#### 14.3.2.1 Reclassification

This involves the process of reassigning a value, a range of values, or a list of values in a raster to new output values, in a new output raster. If one class (or group or range of classes) is more interesting to us than the other classes, its original values can be assigned a specific value and all the others can be changed into a different (background) value. This involves the creation of a discrete raster from either a continuous one or another discrete raster. Reclassification can be applied to both vector and raster objects.

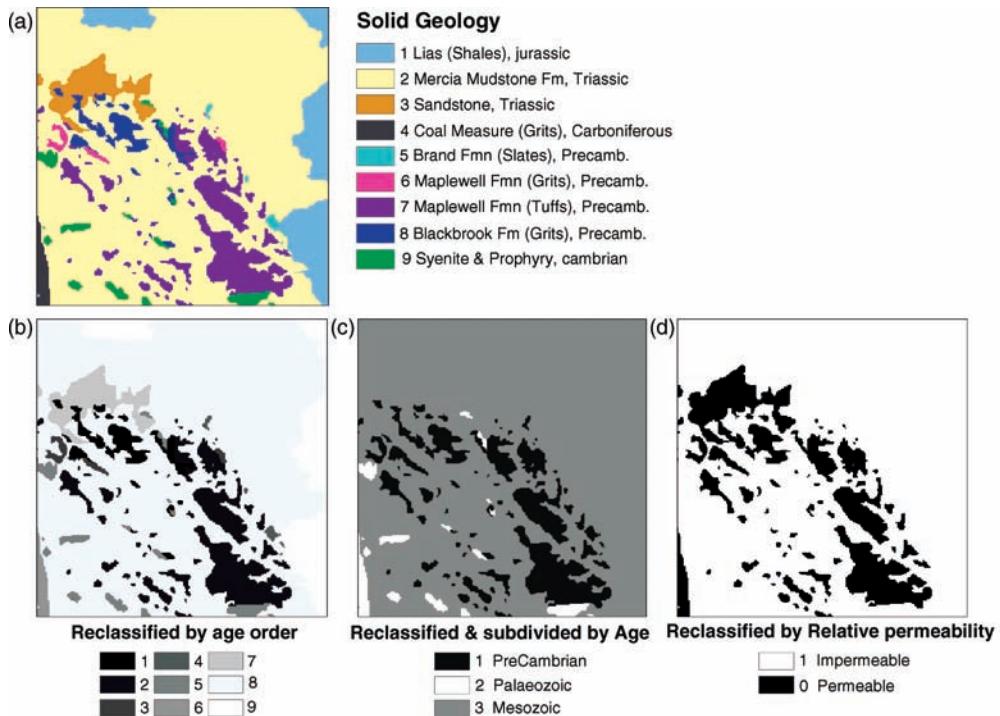
In the case of discrete raster data, a reclassification may be required to produce consistent units among a set of input raster images, in which case a one-to-one value change may be applied. The output raster would look no different, spatially, from the input, having the same number of classes, but the values would have changed.

Different classes or types of feature may be reclassified according to some criteria that are important to the overall analysis. During the reclassification process, weighting can be applied to the output values to give additional emphasis to the significant classes, and at the same time reducing the significance of other classes.

The example in Figure 14.5a shows a discrete raster representation of a geological map in which nine lithological units are coded with values 1 to 9 and labelled for the purposes of presentation,

according to their name, rock type and ages. For the purposes of some analysis it may be necessary to simplify this lithological information, for example according to the broad ages of the units, PreCambrian, Palaeozoic and Mesozoic, for instance. The result of such a simplification is shown in Figure 14.5c; now the map has only three classes and it can be seen that the older rocks (Precambrian and Palaeozoic) are clustered in the south-western part of the area, with the younger rocks (Mesozoic) forming the majority of the area as an envelope around the older rocks. So the simplification of the seemingly quite complex lithological information shown in Figure 14.5a has revealed spatial patterns in that information which are of significance and which were not immediately apparent beforehand. Figure 14.5d shows a second reclassification of the original lithological map, this time on the basis of relative permeability. The information is again simplified by reducing the number of classes to two, impermeable and permeable. Such a map might form a useful intermediary layer in an exercise to select land suitable for waste disposal but also illustrates that subjective judgements are involved at the early stage of data preparation. In the very act of simplifying information, we introduce bias and, strictly speaking, error into the analysis. We also have to accept the assumptions that the original classes are homogeneous and true representations everywhere on the map, which they may not be. In reality there is almost certainly heterogeneity within classes and the boundaries between the classes may not actually be as rigid as our classified map suggests (these matters are discussed further in Chapter 17).

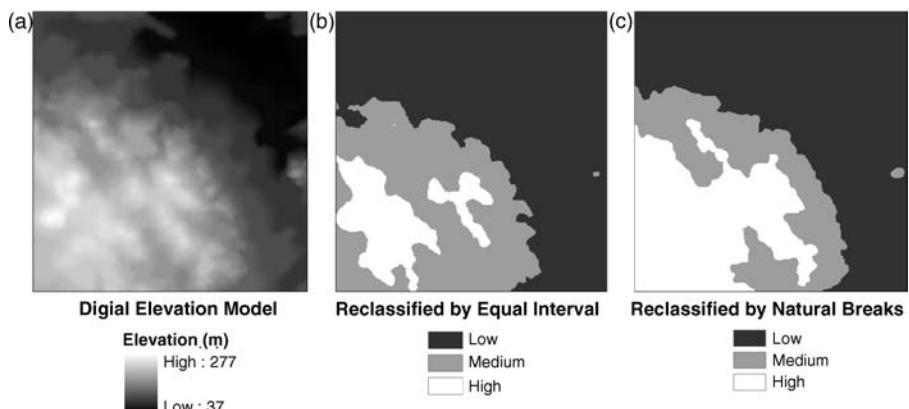
Continuous raster data can also be reclassified in the same way. The image in Figure 14.6a shows a DEM of the same area with values ranging between 37 and 277, representing elevation in metres above sea level. Reclassification of this dataset into three classes of equal interval to show areas of low, medium and high altitude produces the simplified image in Figure 14.6b. Comparison with Figure 14.5b shows that the areas of high elevation coincide with the areas where older rocks exist at the surface in the south-west of the area, again revealing spatial patterns not immediately evident in the original image. Reclassification of the DEM into three classes, this time with the classes defined



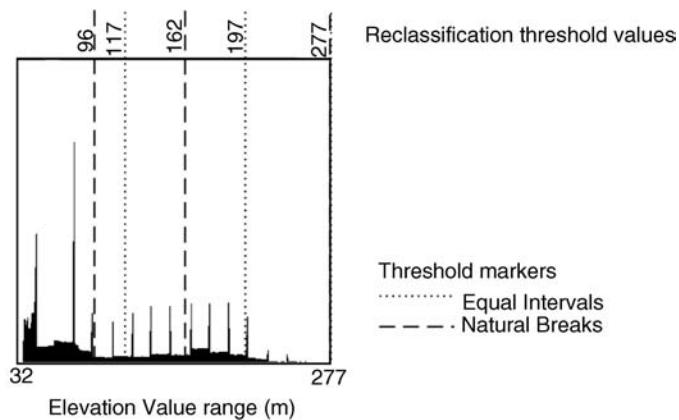
**Figure 14.5** (a) Discrete raster representation of a geological map, with nine classes representing different lithologies; (b) one-to-one reclassification by age order (1 representing the oldest, 9 the youngest); (c) a reclassified and simplified version where the lithological classes have been grouped and recoded into three broad age categories (PreCambrian, Palaeozoic and Mesozoic); (d) a second reclassified version where the lithologies have been grouped according to their relative permeability, with 1 representing impermeable rocks and 0 permeable; such an image could be used as a mask

according to the natural breaks in the image histogram (shown in Figure 14.7), produces a slightly different result, Figure 14.6c. The high-elevation areas are again in the south-west but the shape and

distribution of those areas are different. This demonstrates several things. Firstly, that very different results can be produced when we simplify data so that (and secondly) we should be careful in doing so,



**Figure 14.6** (a) A DEM; (b) a DEM reclassified into three equal interval classes; and (c) a DEM reclassified into three classes by natural breaks (shown in Figure 14.7)



**Figure 14.7** Image histogram of the DEM shown in Figure 14.6a and the positions of the reclassification thresholds set by equal interval and natural break methods (shown in Figure 14.6b and c, respectively)

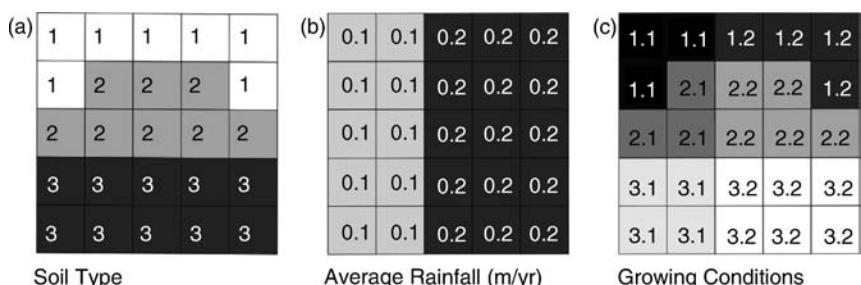
and, thirdly, that the use of the image histogram is fundamental to the understanding of and sensible use of reclassification of continuous raster data. This issue is revisited later in Chapter 15.

Reclassification forms a very basic but important part of spatial analysis, in the preparation of data layers for combination, in the simplification of layer information and especially when the layers have dissimilar value ranges. Reclassification is one of several methods of producing a common range among input data layers that hold values on different measurement scales.

Clear examples of the use of reclassification within case studies can be found in Sections 20.3 and 21.2.

### 14.3.3 Binary operations

This description refers to operations in which there are two input layers, leading to the production of a single output layer. *Overlay* refers to the combination of more than one layer of data, to create one new layer (using the standard operators described in Section 14.2.4). The example shown in Figure 14.8 illustrates how a layer representing average rainfall, and another representing soil type, can be combined to produce a simple, qualitative map showing optimum growing conditions for a particular crop. Such operations are equivalent to the application of formulae to multi-band images, to generate ratios, differences and other



**Figure 14.8** An example of a simple overlay operation involving two input rasters: (a) an integer raster representing soil classes (class 2, representing sandy loam, is considered optimum); (b) a floating-point raster representing average rainfall, in metres per year (0.2 is considered optimum); and (c) the output raster derived by addition of a and b to produce a result representing conditions for a crop; a value of 2.2 (2 + 0.2), on this rather arbitrary scale, represents optimum growing conditions and it can be seen that there are five pixel positions which satisfy this condition

inter-band indices (as described in Chapter 3), and as mentioned in relation to point operations on multi-spectral images, it is important to consider the value ranges of the input bands or layers, when combining their values arithmetically in some way. Just as image differencing requires some form of stretch applied to each input layer, to ensure that the real meaning of the differencing process is revealed in the output, here we should do the same. Either the inputs must be scaled to the same value range, or if the inputs represent values on an absolute measurement scale then those scales should have the same units.

The example shown in Figure 14.8 represents two inputs with relative values on arbitrary nominal or ordinal (Figure 14.8a) and interval (Figure 14.8b) scales. The resultant values are also given on an interval scale and this is acceptable providing the range of potential output values is understood, having first understood the value ranges of the inputs, since they may mean nothing outside the scope of this simple exercise.

Another example could be the combination of two rasters as part of a *cost-weighted analysis* and possibly as part of a wider *least cost pathway* exercise. The two input rasters may represent measures of cost, as produced through reclassification of, for instance, slope angle and land value, cost here being a measure of friction or the real cost of moving or operating across the area in question. These two cost rasters are then *aggregated* or *summed* to produce an output representing total cost for a particular area (Figure 14.9).

### 14.3.4 N-ary operations

Here we deal with a potentially unlimited number of input layers to derive any of a series of standard statistical parameters, such as the mean, standard deviation, majority and variety. Ideally there should be a minimum of three layers involved but in many instances it is possible for the processes to be performed on single layers; the result may, however, be rather meaningless in that case. The more commonly used statistical operations and their functionalities are summarized in Table 14.4. As with the other local operations, these statistical parameters are point operations derived for each individual pixel position, from the values at corresponding pixel positions in all the layers, rather than from the values within each layer (as described in Chapter 2).

## 14.4 Neighbourhood operations

### 14.4.1 Local neighbourhood

These can be described as being incremental in their behaviour or operation. They work within a small neighbourhood of pixels (which in some circumstances can be user defined) to change the value of the pixel at the centre of that neighbourhood, based on the local neighbourhood statistics. The process is then repeated, or incremented, to the next pixel position along the row, and so on until the

(a)	<table border="1"> <tr><td>12</td><td>17</td><td>24</td><td>28</td><td>30</td></tr> <tr><td>13</td><td>15</td><td>22</td><td>27</td><td>29</td></tr> <tr><td>9</td><td>11</td><td>18</td><td>24</td><td>27</td></tr> <tr><td>7</td><td>8</td><td>12</td><td>16</td><td>21</td></tr> <tr><td>3</td><td>4</td><td>7</td><td>14</td><td>16</td></tr> </table>	12	17	24	28	30	13	15	22	27	29	9	11	18	24	27	7	8	12	16	21	3	4	7	14	16	(b)	<table border="1"> <tr><td>6</td><td>7</td><td>8</td><td>9</td><td>9</td></tr> <tr><td>6</td><td>7</td><td>8</td><td>9</td><td>9</td></tr> <tr><td>5</td><td>6</td><td>7</td><td>8</td><td>9</td></tr> <tr><td>5</td><td>5</td><td>6</td><td>7</td><td>8</td></tr> <tr><td>4</td><td>4</td><td>5</td><td>6</td><td>7</td></tr> </table>	6	7	8	9	9	6	7	8	9	9	5	6	7	8	9	5	5	6	7	8	4	4	5	6	7	(c)	<table border="1"> <tr><td>3</td><td>3</td><td>2</td><td>1</td><td>1</td></tr> <tr><td>5</td><td>6</td><td>6</td><td>5</td><td>1</td></tr> <tr><td>6</td><td>5</td><td>9</td><td>9</td><td>6</td></tr> <tr><td>7</td><td>8</td><td>9</td><td>9</td><td>8</td></tr> <tr><td>6</td><td>8</td><td>9</td><td>8</td><td>7</td></tr> </table>	3	3	2	1	1	5	6	6	5	1	6	5	9	9	6	7	8	9	9	8	6	8	9	8	7	(d)	<table border="1"> <tr><td>9</td><td>10</td><td>10</td><td>10</td><td>10</td></tr> <tr><td>11</td><td>13</td><td>14</td><td>14</td><td>10</td></tr> <tr><td>11</td><td>11</td><td>16</td><td>17</td><td>15</td></tr> <tr><td>12</td><td>13</td><td>15</td><td>16</td><td>16</td></tr> <tr><td>10</td><td>12</td><td>14</td><td>14</td><td>14</td></tr> </table>	9	10	10	10	10	11	13	14	14	10	11	11	16	17	15	12	13	15	16	16	10	12	14	14	14
12	17	24	28	30																																																																																																							
13	15	22	27	29																																																																																																							
9	11	18	24	27																																																																																																							
7	8	12	16	21																																																																																																							
3	4	7	14	16																																																																																																							
6	7	8	9	9																																																																																																							
6	7	8	9	9																																																																																																							
5	6	7	8	9																																																																																																							
5	5	6	7	8																																																																																																							
4	4	5	6	7																																																																																																							
3	3	2	1	1																																																																																																							
5	6	6	5	1																																																																																																							
6	5	9	9	6																																																																																																							
7	8	9	9	8																																																																																																							
6	8	9	8	7																																																																																																							
9	10	10	10	10																																																																																																							
11	13	14	14	10																																																																																																							
11	11	16	17	15																																																																																																							
12	13	15	16	16																																																																																																							
10	12	14	14	14																																																																																																							
Slope gradient (deg)		Ranked slope (friction 1)		Ranked value (friction 2)																																																																																																							
Total cost = f1 + f2																																																																																																											

**Figure 14.9** (a) Slope gradient in degrees; (b) ranked (reclassified) slope gradient constituting the first cost or friction input; (c) ranked land value (produced from a separate input land-use raster) representing the second cost or friction input; and (d) total cost raster produced by aggregation of the input friction rasters ( $f_1$  and  $f_2$ ). This total cost raster could then be used within a cost-weighted distance analysis exercise

**Table 14.4** Summary of local pixel statistical operations, their functionality and input/output data format

Statistic	Input format	Functionality	Data type
Variety	Only rasters. If a number is input, it will be converted to a raster constant for that value	Reports the number of different DN values occurring in the input rasters	Output is integer
Mean		Reports the average DN value among the input rasters	Output is floating point
Standard deviation	Rasters, numbers and constants	Reports the standard deviation of the DN values among the input rasters	Output is floating point
Median		Reports the middle DN value among the input raster pixel values. With an even number of inputs, the values are ranked and the middle two values are averaged. If inputs are all integer, output will be truncated to integer	
Sum		Reports the total DN value among the input rasters	
Range		Reports the difference between maximum and minimum DN value among the input rasters	
Maximum		Reports the highest DN value among the input rasters	
Minimum	Only rasters. If a number is input, it will be converted to a raster constant for that value	Reports the lowest DN value among the input rasters	If inputs are all integer, output will be integer, unless one is a float, then the output will be a float
Majority		Reports the DN value which occurs most frequently among the input rasters. If no clear majority, output = null, for example if there are three inputs all with different values. If all inputs have equal value, output = input	
Minority		Reports the DN value which occurs least frequently among the input rasters. If no clear minority, as majority If only two inputs, where different, output = null. If all inputs equal, output = input. If only one input, output = input	

whole raster has been processed. It is equivalent to convolution filtering in the image or spatial domain, as described in some detail in Section 4.2. In image processing the process is used to quantify or enhance the spatial patterns or textures of a remotely

sensed image, for instance. Here we are often dealing with data that are of implied 3D character, for example the gradient or curvature (Laplacian) of surface topography (see also Chapter 16) and, if so, we are using the same process to quantify, describe

or extract information relating to the morphology of the surface described by the DN in the local neighbourhood. Examples include calculations of slope angle, aspect and curvature, and mathematical morphology such as collapsing and expanding raster regions. These are described in more detail in Chapter 16. Otherwise these neighbourhood processes can be used to simplify or generalize discrete rasters.

#### 14.4.1.1 Distance

Mapping distance allows the calculation of the proximity of any raster pixel to/from a set of target pixels, to determine the nearest or to gain a measure of cost in terms of distance. This is classified as neighbourhood processing since the value assigned to the output pixel is a function of its position in relation to another pixel. The input is a discrete raster image, in which the target pixels are coded, probably with a value of 1 against a background of 0 (as illustrated in Figure 14.10a). This input image may in itself be the product of an earlier reclassification. The simplest form of this operation involves the use of a *straight line distance* function, which calculates the *Euclidean distance* from every pixel to the target pixels (Figure 14.10b). Most GISs will also offer a spherical Earth calculation as an alternative which does not use any georeferencing (projection) information.

The output pixel values represent the Euclidean distance from the target pixel centres to every other pixel centre and are coded in the value units of the input raster, usually metres, so that the input raster will usually contain integers and the output

normally floating-point numbers. The calculated distance raster may then be further *reclassified* for use as input to more complex multi-criteria analysis or used within a cost-weighted distance analysis.

#### 14.4.1.2 Cost pathways

This moving window or kernel procedure is used to derive a cost-weighted distance and cost-weighted direction (these are referred to slightly differently depending on which software product you are using) as part of a least cost pathway exercise. The cost-weighted distance function operates by evaluating each input pixel value of a *total cost* raster (as in figure 14.9) and comparing it with its neighbouring pixels. The average cost between each is multiplied by the distance between them. Cost-weighted direction is generated also from the total cost raster, where each pixel is given a value using a direction-encoded  $3 \times 3$  kernel, which indicates the direction to the lowest cost pixel value from among its local neighbours. These two rasters or surfaces are then combined to derive the least cost pathway or route across the raster, to the target.

#### 14.4.1.3 Mathematical morphology

Mathematical morphology can be thought of as the combination of map algebra and set theory, or of conditional processing and convolution filtering. As a concept it was first developed by Matheron (1975) and then subsequently by many others. It describes the spatial expansion and shrinking of objects through neighbourhood processing and extends the concept of filtering. Such changes include *erosion* or *shrinking*, *dilation* or *expansion*, *opening* and *closing* of raster images. The size and shape of the neighbourhoods used are controlled by *structuring elements* or *kernels* which may be of varying size and form. At its simplest, a kernel is a set of values passed across a binary raster image, whose status (1 or zero, ‘on’ or ‘off’) is changed according to agreement with the values in the kernel. The processing may not be reversible; for instance, after eroding such an image, using an erosion kernel, it is generally not possible to return the binary image to its original shape through the dilation kernel. Several different kinds of structuring kernels can be used, including those which are square, in addition to 1D, hexagonal, circular and irregularly shaped ones.

(a)	0	0	0	0	0
Target raster	0	0	0	0	0
	0	0	0	0	0
	0	1	0	0	0
	0	0	0	0	0

(b)	25.4	32	25.4	28.9	33.9
Ditance raster	18.5	16	18.5	22.6	28.9
	11.3	8	11.3	18.5	25.4
	8	0	8	16	32
	11.3	8	11.3	18.5	25.4

**Figure 14.10** (a) An input discrete (binary) raster and (b) the straight line or Euclidean distance calculated from a single target or several targets are coded to every other pixel in an input

Mathematical morphology can be applied to vector point, line and area features but more often involves raster data, commonly discrete rasters and sometimes continuous raster surfaces, such as DEMs. In the last case it can be used to find and correct for errors or extreme values (high or low) in those surfaces. Here we concentrate on the mechanism of the operations involved. It has also been used in mineral prospection mapping, to generate evidence maps, and in the processing of rock thin-section images, to find and extract mineral grain boundaries. It is a method which has applications in raster topology and networks, in addition to pattern recognition, image texture analysis and terrain analysis. These and related methods have also been developed for edge feature extraction and image segmentation, for example the Canny edge detector and OCR text recognition (Parker, 1997).

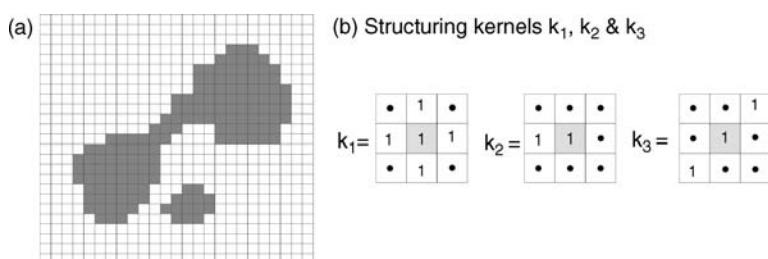
To illustrate the effects, consider a simple binary raster image showing two classes, as illustrated in Figure 14.11. The values in the raster of the two classes are 1 (inner, dark grey class) and 0 (surrounding, white class); the image consists of a grid of ones and zeros. This input raster is processed using a series of  $3 \times 3$  structuring elements or kernels ( $k$ ), which consist of the values 1 and null (rather than 0).

The kernels are passed incrementally over the raster image, changing the central pixel each time, according to the pattern of its neighbouring values. The incremental neighbourhood operation is therefore similar to spatial filtering but with conditional rather than arithmetic rules controlling the modification of the central value.

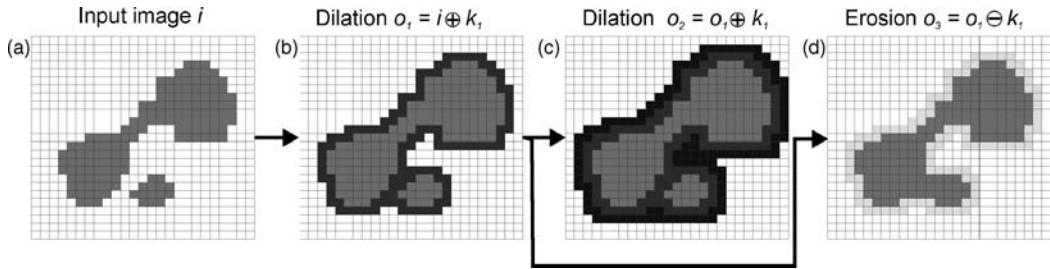
A simple dilation operation involves the growth, or expansion, of an object and can be described by

$$o = i \oplus k \text{ or } \delta k(i) \quad (14.1)$$

where  $o$  is the output binary raster,  $i$  is the input binary raster and  $k$  is the kernel which is centred on a pixel at  $i$ , and  $\delta$  indicates a dilation. The Minkowski summation of sets ( $a \oplus b$ ) refers to all the pixels in  $a$  and  $b$ , in which  $\oplus$  is the vector sum, and  $a$  belongs to set  $b$ , and  $b$  belongs to set  $a$  (Minkowski, 1911); the Minkowski effect is where one shape is grown by the shape of another. The values of  $i$  are compared with the corresponding values in the kernel  $k$ , and are modified as follows: the value in  $o$  is assigned a value of 1 if the central value of  $i$  equals 1, or if any of the other values in  $k$  match their corresponding values in  $i$ ; if they differ, the resultant value in  $o$  will be 0. The result of this is to leave the inner values as they are and to modify the surrounding outer values by the morphology of the kernel. The effect of a dilation, using kernel  $k_1$ , is to add a rim of black pixels around the inner shapes, and in doing so the two shapes in the binary image are joined into one, both having been dilated, as in Figure 14.12b. If the output  $o_1$  is then dilated again using  $k_1$ , then a second rim of pixels is added, and so on. It can be seen that by this process, the features are merged into one. Using these conditional rules, the effect of a dilation can be considered equivalent to a maximum operation. Dilation is commonly used to create a map or image that reflects proximity to or distance from a feature or object, such as distance from road networks or proximity to major faults. These distance or ‘buffer’ maps often form an important part of multi-layer spatial analysis, such as in the modelling of mineral prospecting, where proximity to a particular phenomenon is considered a significant and favourable condition.



**Figure 14.11** (a) Simple binary raster image ( $i$ ); and (b) the three structuring kernels ( $k_1$ ,  $k_2$  and  $k_3$ ) the effects of which are illustrated in Figures 14.12–14.14. The black dots in the kernels represent null values



**Figure 14.12** Dilation, erosion and closing: (a) the original image ( $i$ ); (b) dilation of  $i$  using  $k_1$  to produce  $o_1$ ; (c) dilation of  $o_1$  also using  $k_1$  to produce  $o_2$ ; and (d) erosion of  $o_1$  using  $k_1$  to produce  $o_3$ . Notice that  $o_3$  cannot be derived from  $i$  by a simple dilation using  $k_1$ ; the two objects are joined and this effect is referred to as closing. The pixels added by dilation are shown black and those pixels lost through erosion are shown with pale grey tones

A simple erosion operation ( $a \ominus b$ ) has the opposite effect, where  $\ominus$  is a vector subtraction, so that it involves the shrinking of an object using the Minkowski subtraction, and is described by

$$o = i \ominus k \text{ or } \varepsilon k(i) \quad (14.2)$$

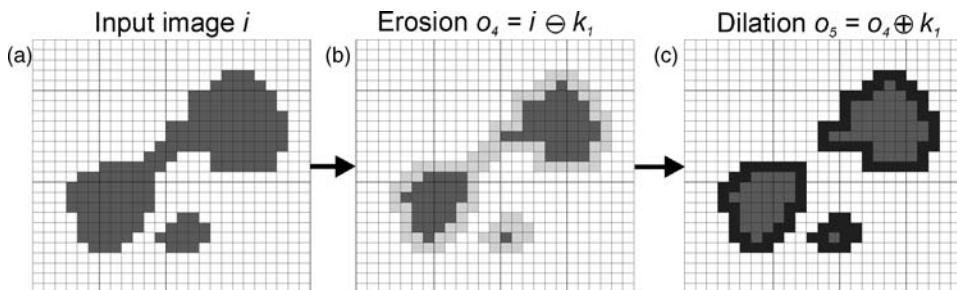
where  $\varepsilon$  indicates an erosion. The values in  $o$  are compared with those in  $k$  and if they are the same then the pixel is ‘turned off’, i.e. the value in  $o$  will be set to 0, or left unchanged if they are not the same. The effect of this, using kernel  $k_1$ , is the removal of a rim of value 1 (grey) pixels from the edges of the feature shown in Figure 14.12b to produce that shown in Figure 14.12d. Using these conditional rules, the effect of an erosion operation can be considered equivalent to a minimum operation. Notice that the output,  $o_3$ , which is the product of the sequential dilation of  $i$ , then erosion of  $o_1$ , results in the amalgamation of the two original objects, and that the subsequent shrinking produces a general-

ized object which covers approximately the area of the original, an effect known as *closing* (dilation followed by erosion). Notice also that the repeated erosion of  $o_3$  will not restore the appearance of the two original features in  $i$ .

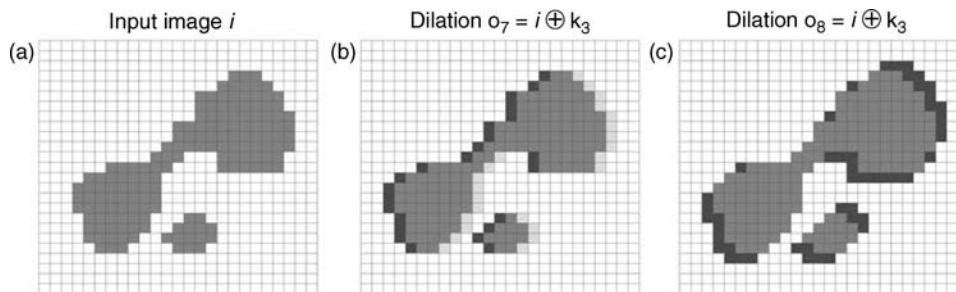
In Figure 14.13b, an erosion operation is performed on the original  $i$ , removing one rim of pixels, and causes the feature to be subdivided into two. When this is followed by a dilation, the result is to restore the two features to more or less their original size and shape except that the main feature has been split into two. This splitting is known as an *opening* (erosion followed by dilation) and is shown in Figure 14.13c:

$$\begin{aligned} \text{Opening, } \gamma k(i) &= \delta k[\varepsilon k(i)] \\ \text{Closing, } \phi k(i) &= \varepsilon k[\delta k(i)]. \end{aligned} \quad (14.3)$$

Again, repeated dilations of the features after opening will not restore the features to their appearance in  $i$ .



**Figure 14.13** Erosion, dilation and opening: (a) the original image ( $i$ ); (b) erosion of  $i$  using  $k_1$  to produce  $o_4$ ; (c) subsequent dilation of  $o_4$ , using  $k_1$ , to produce  $o_5$ . Note that the initial erosion splits the main object into two smaller ones and that the subsequent dilation does not restore the object to its original shape, an effect referred to as opening



**Figure 14.14** Anisotropic effects: (a) the original image ( $i$ ); (b) dilation of  $i$  using  $k_2$  to produce  $o_6$ , causing a westward shift of the object; and (c) dilation of  $i$  using  $k_3$ , producing an elongation in the NE-SW directions to produce  $o_7$

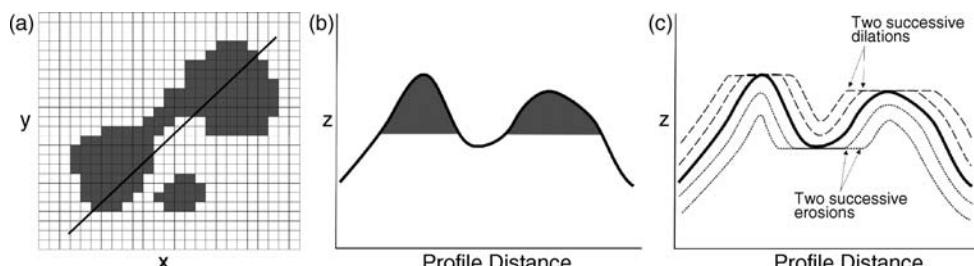
Closing can be used to generalize objects and to reduce the complexity of features in a raster, such as where a cluster of small features all representing the same class are dilated into one region representing that class and then eroded to reduce the features to approximately the same area as before but with reduced complexity. Opening can be used to perform a kind of sharpening or to add detail or complexity to the image.

Dilation and erosion operations can also be carried out anisotropically, i.e. they can be applied by unequal amounts and in specific directions. Such directional operations are often relevant in geological applications where there is some kind of structural or directional control on the phenomenon of interest.

For example, the effect of kernel  $k_2$  on  $i$  is shown in Figure 14.14a, where the effect is a westward shift of the features by 1 pixel. The effect of kernel  $k_3$  is to cause dilation in the NW–SE directions, resulting in an elongation of the feature (Figure 14.14b).

To consider the effect of mathematical morphology on continuous raster data, we can simply take the binary image ( $i$ ) shown in Figures 14.11–14.14

to represent a density slice through a raster surface, such as an elevation model. In this case, the darker class would represent the geographical extent of areas exceeding a certain elevation value. Figure 14.15a shows the binary image and a line of profile (Figure 14.15b) across a theoretical surface which could be represented by image ( $i$ ). The effect of simple dilation and erosion of the surface is shown in Figure 14.15c; it can be seen that dilations would have the effect of filling pits or holes, and broaden peaks in the surface, while erosions reduce the peaks or spikes, and widen depressions. Such techniques could therefore be used to correct for errors in generated surfaces, such as DEMs, except that the dilations and erosions affect all other areas of the surface too, including the parts which do not need correcting. Such artefacts and errors in DEMs cannot be properly corrected by merely smoothing either, since the entire DEM will also be smoothed and so degraded. The use of median filters to smooth while retaining edge features has been proposed but, again, this is also undesirable for the same reason. A modification of the mathematical morphology



**Figure 14.15** (a) The original input image with the position of a profile line marked; (b) the theoretical cross-sectional profile with the shaded area representing the geographical extent of the darker class along the line shown in (a); and (c) the effect on the profile of dilations and erosions of that surface

technique has been proposed, known as *morphological reconstruction* (Vincent, 1993), for the correction of DEM errors. In this case, the original image is used as a *mask* and the dilations and erosions are performed iteratively on a second version of the same image (*marker image*) until stability between the mask and marker images is reached and the image is fully reconstructed and no longer changes, when the holes or pits are corrected (Figure 14.16). Since morphological reconstruction is based on repeated dilations, rather than directly modifying the surface morphology, it works by controlling connectivity between areas. The marker could simply be created by making a copy of the mask and either subtracting or adding a constant value. The error-affected raster image is then used as the mask, and the marker (which is derived from it) is dilated or eroded repeatedly until it is constrained by the mask, i.e. until there is no change between the two, and the process then stops. By subtracting a constant from the marker and repeatedly dilating it, extreme peaks can be removed, whereas by adding a constant and repeatedly eroding the marker, extreme pits would be removed. The extreme values are effectively reduced in magnitude, relative to the entire image value range, in the reconstructed marker image. This technique can be (and has been) used selectively to remove undesirable extreme values from DEMs.

#### 14.4.2 Extended neighbourhood

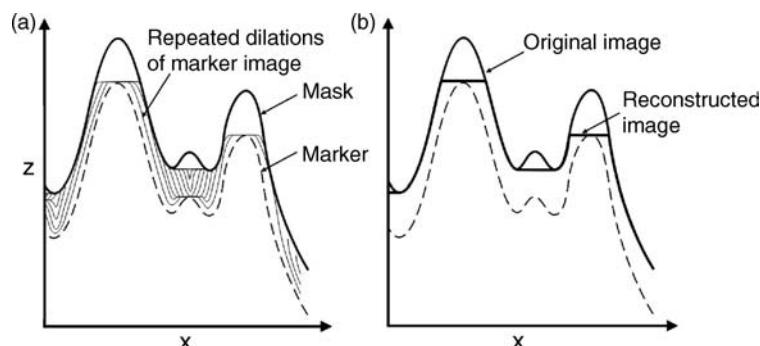
The term ‘extended neighbourhood’ is used to describe operations whose effects are constrained

by the geometry of a feature in a layer and performed on the attributes of another layer. These extended neighbourhood operations can be further described as focal and zonal. If for instance slope angles must be extracted from within a corridor along a road or river, the corridor is defined from one layer and then used to constrain the extent of the DEM from which the slope angle is then calculated (see Figure 14.17).

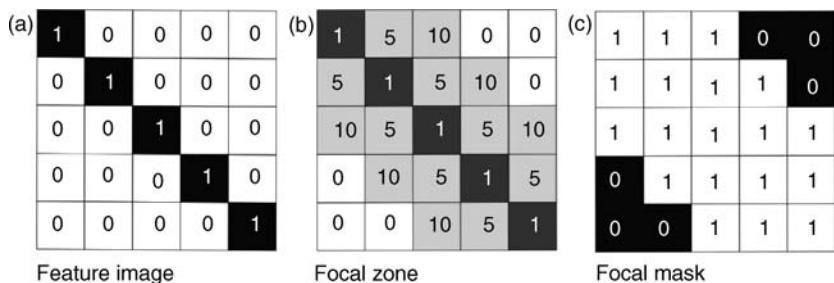
##### 14.4.2.1 Focal operations

A *focal operation* is used for generating corridors and buffers around features. Focal operations are those that derive each new value as a function of the existing values, distances and/or directions of neighbouring (but not necessarily adjacent) locations. The relationships may be defined by such variables as Euclidean distance, travel cost, engineering cost or inter-visibility. Such operations could involve measurement of the distance between each pixel (or point) position and a target feature(s). A buffer can then be created by reclassification of the output ‘distance’ layer.

This allows specific values to be set for the original target features, with the buffer zones and for the areas beyond the buffers. In this way, it is possible to establish the approximate proximity of objects using a buffer. Buffer zones can also be used as masks to identify all features that lie within a particular distance of another feature. Buffers can be set at a specified distance or at a distance set by an attribute. Since the buffer is a reclassification of the distance parameter, multiple buffer rings can also be easily generated. Buffers are therefore particu-



**Figure 14.16** Mechanism of morphological reconstruction of an image, as illustrated by a profile across the image: (a) in this case, by repeated dilations of the marker until it is constrained by the mask image; (b) the extreme peaks are reduced in magnitude in the reconstructed image



**Figure 14.17** Focal statistics: (a) a binary image representing a linear target feature (coded with a value of 1 for the feature and 0 for the background); (b) a 10 m focal image created around the linear feature, where each pixel is coded with a value representing its distance from the feature (assuming that the pixel size is 5 m × 5 m), areas beyond 10 m from the feature remain coded as 0; and (c) binary focal zone mask with values of 1 within the mask and zero outside it. This has a similar effect to a dilation followed by a reclassification, to produce a distance buffer

larly useful for constraining the activities of spatial analysis. Dilation, as described previously, is just one method of creating a buffer.

#### 14.4.2.2 Zonal operations

A *zonal operation* also involves the use of the spatial characteristics of a zone or region defined on one layer, to operate on the attribute(s) of a second layer or layers. The zonal areas may be regularly or irregularly shaped. This process falls into the binary operations category since zonal operations most commonly involve two layers. An example is given in Figure 14.18 where zonal statistics are calculated from an input layer representing the density of forest growth, within the spatial limits defined by a second survey boundary layer, to provide an output representing, in this case, the average forest density within each survey

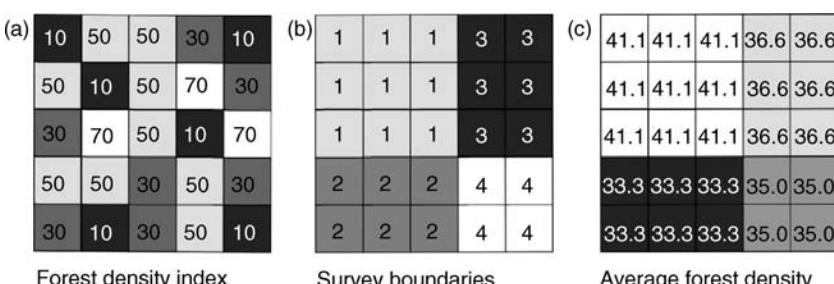
unit. Notice that the two raster inputs contain integer values but that the output values are floating-point numbers, as is always the case with mean calculations.

## 14.5 Vector equivalents to raster map algebra

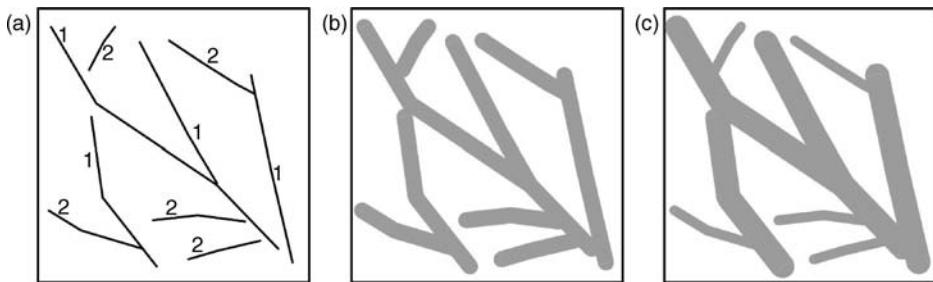
Map algebra operations can be performed on vector data too. The operators behave slightly differently because of the nature of vector data but in many cases are used to achieve the similar results.

### 14.5.1.1 Buffers

A zone calculated as the Euclidean distance from existing vector features, such as roads, is referred to



**Figure 14.18** Zonal statistics: (a) forest density integer image; (b) survey boundaries (integer) image; and (c) the result of zonal statistics (in this case a zonal mean) for the same area. Note that this statistical operation returns a non-integer result



**Figure 14.19** (a) Simple vector line feature map, labelled with attribute values (1 and 2); (b) output with buffers of constant distance; (c) output map with buffers of distance defined by the attribute values shown in (a) (features with attribute value 1 having buffers twice the distance of those with attribute value 2)

as a *buffer*. Buffers are calculated at constant distance from the feature or at distances dictated by attribute values, and each zone will be the same width around the feature (see Figure 14.19). No account is taken of the Earth's curvature, so the zones will be at the same width regardless of the coordinate system. Negative distance values can be used, and these will cause a reduction in the size of the input feature. Buffers can also be generated on only one side of input features (should this be appropriate). The input layer in this case is a vector feature but the output may be a polygon file or raster. The same buffering operation can also be applied to raster data (as described in Section 14.5.2) by first calculating the Euclidean distance and then reclassifying the output to exclude distances within or beyond specified thresholds; the output will always be a raster in this case. Buffering in this way can be considered as the vector equivalent of conditional logic combined with raster dilation or erosion.

#### 14.5.1.2 Dissolve

When boundaries exist between adjacent polygon or line features, they could be removed or *dissolved* because they have the same or similar values for a particular attribute (see Figure 14.20). As in a geological map where adjacent lithological units with similar or identical descriptions can sensibly be joined into one, the boundaries between them are removed by this process and the classes merged into one. Complications in the vector case arise if the features' attribute tables contain other attributes (besides the one of interest being merged) which differ across the boundary; choices must be made about how those other attributes should appear in

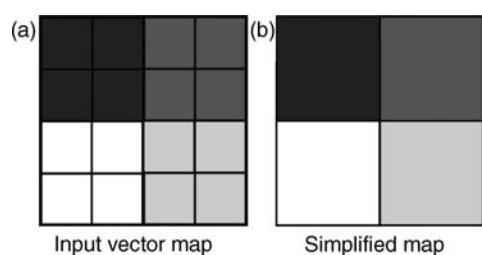
the output dissolved layer. This is equivalent to merging raster classes through reclassification, or raster generalization/simplification.

#### 14.5.1.3 Clipping

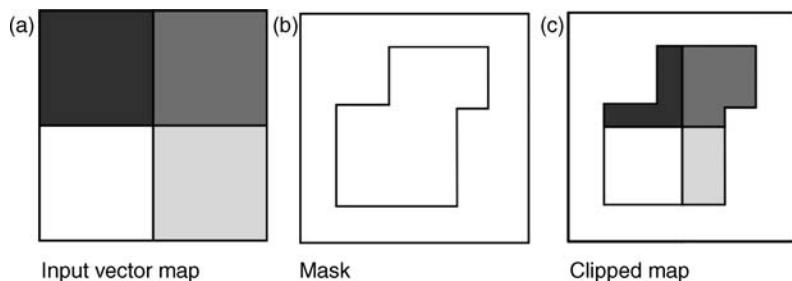
The geometry of a feature layer can be used as a mask to extract selectively a portion of another layer; the input layer is thereby *clipped* to the extent of the mask (see Figure 14.21). The feature layer to be clipped may contain point, line or polygon features but the feature being used as a mask must have area, i.e. it will always be a polygon. The output feature attribute table will contain only the fields and values of the extracted portion of the input vector map, as the attributes of the mask layer are not combined. Clipping is equivalent to a binary raster zonal operation, where the pixels inside or outside the region are set as null, using a second layer to define the region or mask.

#### 14.5.1.4 Intersection

If two feature layers are to be integrated while preserving only those features that lie within the



**Figure 14.20** Vector polygon features (a) and the dissolved and simplified output map (b)



**Figure 14.21** Vector polygon clipping, using an input vector layer from which an area will be extracted (a), the vector feature whose geometric properties will be used as the mask (b) and (c) the output clipped vector feature

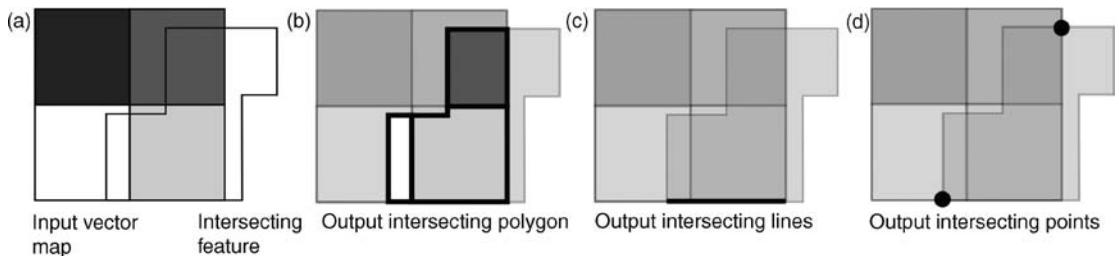
spatial extent of both layers, an *intersection* can be performed (see Figure 14.22). This is similar to the clip operation except that the two input layers are not necessarily of the same feature type. The input layers could be point, line and/or polygon, so the output features could also be point, line and/or polygon in nature. New vertices need to be created to produce the new output polygons, lines and points, through a process called *cracking*. Unlike the clip operation, the output attribute table contains fields and values from both input layers, over the intersecting feature/area. In the case of two intersecting polygons, intersection is equivalent to a Boolean operation using a logical AND (Min) operator between two overlapping raster images.

When two input overlapping feature layers are required to be integrated such that the new output feature layer contains all the geometric features and attributes of two input layers, the *union* operation can be used (see Figure 14.23). Since vector feature layers can contain only points or only polygons, here the inputs must be of the same type but the number of inputs is not limited to two. Again, new

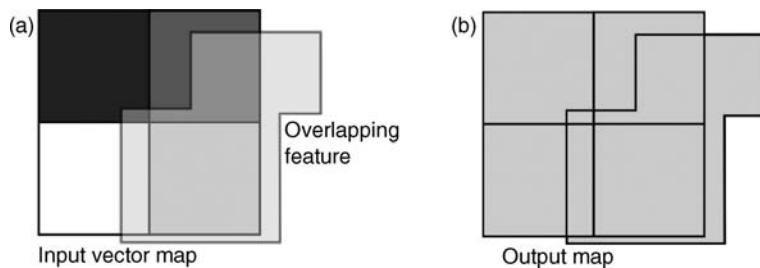
vertices will be created through cracking. This is similar to the intersect operation but the output will have the total extent of the input layers. New, minor polygons are created wherever polygons overlap. The attribute table of the output layer contains attribute fields of both the input layers, though some of the entries may be blank. In the polygon case, it is equivalent to a binary raster operation using a logical OR (Max) operator between overlapping images.

## 14.6 Summary

The overlap with image processing is perhaps most obvious in this chapter. The processing of the DN values in various bands of a multi-spectral image is analogous to that of raster grids in GIS using map algebra. Local, focal, zonal, incremental and global operations in raster GIS are synonymous with those of image processing even though the objective may be different. The use of conditional statements is another parallel and represents the first step in the



**Figure 14.22** Intersection operation between two overlapping polygon features (a); the output intersecting polygon (b) which covers the extent and geometry of the area which the two inputs have in common; the intersecting line (c) and points (d) shared by both polygons. The output attribute table contains only those fields and values that exist over the common area, line and points



**Figure 14.23** Vector polygon union operation where two polygon features overlap (a), and the output object (b) covers the extent and geometry of both inputs. The output attribute table also contains the attribute fields and values of both input features

development of a more complex spatial analysis, decision making. The use of the geometric properties of one layer to control the limits of operations on another is a minor departure since this is less commonly required in image processing but it is perfectly possible using ‘regions of interest’ for which statistics have been derived (Chapter 3). Regions are spatially defined within the coordinate space of the image, the extents of which are recorded in association with the raster image information (header). Statistics can be calculated globally and for the region and these can be manipulated on any of the bands of the image. These are in essence zonal operations.

This chapter focused on operations that assume raster inputs. Many of these operations have vector equivalents, and in some circumstances it could be argued that they could be carried out more effectively using vector data. The diversity of raster-based operations is, however, testament to their significance in the processing of continuously sampled data whose spatial variability is significant. This is especially the case in geoscientific applications, where we are deeply concerned with the way variables change from one location to another as well as the spatial relationships between the variables.

Map algebra plays a major role in multi-criteria and multi-objective problems, by linking together these simple processes and procedures to prepare

data and build complex models, and so to tackle complex spatial problems, which are discussed further in Chapter 18.

## Questions

- 14.1 With respect to the nature of the classes being represented, what assumptions are made during spatial operations on categorical (discrete) rasters?
- 14.2 How should we use these tools to scale (prepare) our data as input for spatial analysis? And what are the effects of using those scales?
- 14.3 How should you decide on the threshold values for reclassification schemes?
- 14.4 Why is it important to understand the nature of the input recorded data when applying local statistical operations?
- 14.5 What are the practical applications of mathematical morphology?
- 14.6 For further consideration beyond this chapter (see Chapter 18):
  - (a) How do these individual operations combine and contribute to more complex spatial analytical models?
  - (b) Are discrete and continuous rasters treated differently within spatial analyses?



# 15

## Extracting Information from Point Data: Geostatistics

### 15.1 Introduction

The data that we have at our disposal are never complete; we have either the wrong kind or insufficient or partial coverage. Naturally, we seek ways to predict the values between, or to extrapolate beyond, the limits of our data; indeed, therein also lies the role of multi-criteria spatial analysis, but we will deal with that in Chapter 18.

This chapter deals with two topics: gaining a better understanding of the data and dealing with incomplete data. If we understand the nature and meaning of the sample data that we do have, we will have a better chance of producing a reliable prediction of the unknowns. After all, one of the most important messages of this book is that producing an impressive result is not enough; if it cannot be explained or understood, it is meaningless.

This chapter therefore covers the subject of *geostatistics*, a term first coined with ‘trepidation’ by Hart (1954) and first used in mineral resource evaluation in an attempt to predict the potential economic value of a mineral deposit from limited sample data by George Matheron and Daniel Krige in 1951. Such techniques have subsequently been applied to many disciplines other than the geosciences. The many and varied uses of geostatistics include, for example, the description and summary

of spatial data attributes, simplifying complex spatial patterns, inferring the characteristics of a larger population on the basis of a sample, estimating the probability of an outcome location and establishing how closely a predicted spatial pattern matches an actual one. Geostatistics is concerned with the description of patterns in spatial data; each known data point has a geographic location and a value, and the connection between them is exploited to help predict values at the unknown locations. There are many, truly comprehensive accounts of geostatistical methods which are listed in the general references and further reading section. We aim only to give an overview of the main issues and methods involved in extracting and exploiting statistical data, and in getting over the problem of incomplete data.

Early qualitative questions about the nature of processes and phenomena have quickly developed into more quantitative questions, ‘how much’ or ‘to what degree’ and ‘how sure are we that the result is true or representative?’ This touches on the issue of uncertainty in data and analysis (this is discussed in more detail in the Chapter 17). Asking ‘why’ is rather more tricky for GIS to tackle since it requires the unravelling of causative links between phenomena and this is a dangerously speculative area.

In dealing with the estimation of unknown values from known ones, this chapter also overlaps with topics in Section 9.2.

## 15.2 Understanding the data

We should never underestimate the importance of understanding our data, how they were collected, how reliable and accurate their geographic positions are, what area of ground they represent and whether their values represent one or more statistically independent populations; all these should be considered when thinking about how to process data and interpret the result.

### 15.2.1 Histograms

A vital tool in the understanding of the data, and one which should be our first port of call, is the histogram. It shows us the count of data points falling into various ranges (the frequency distribution). The histogram shows us the general shape and spread, symmetry of distribution and modality of the data, and should reveal outliers.

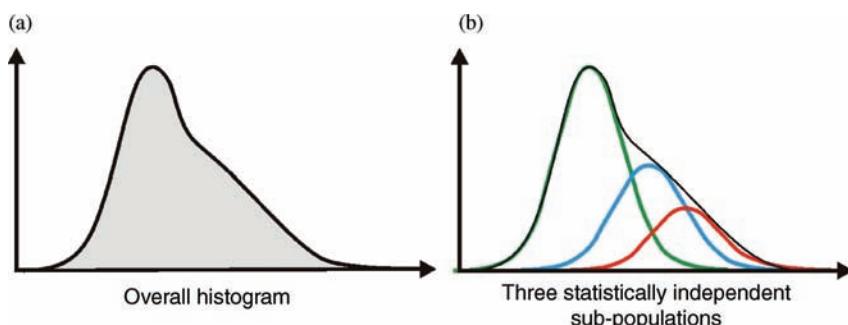
No interpolation method should be used without a full understanding of the implications and effects on the result. If any of the methods that are used to interpolate data comprises more than one statistically independent population, then the result is flawed. Or, if predictions are to be based on calculations from such datasets which contain more than one population, the actual recorded values could

be considerably less than the predicted ones. For instance, in the case of predicted output of metal from an actively producing mine, predictions of total metal production based on the concentrations measured at discrete sample points within the mine may significantly overestimate production if the existence of different and unrelated mineralisations within the sample data is not recognized.

Here again the data histogram becomes a vital tool in understanding the data. It should always be carefully examined beforehand, to establish how the data are composed. If the data are not normally distributed about the mean but skewed, there must be a reason for it, and it will be necessary to consider the existence of minor populations as the reason (see Figure 15.1). Skewness in the data histogram could indicate a sampling, measurement or processing problem, or it could point to some real but unknown pattern in the data; either way, it should be investigated before proceeding. There are two important messages here: firstly, the possibility of erroneous numerical predictions from the data and, secondly, the fact that interpolating such data across multiple populations could produce a surface which is meaningless.

### 15.2.2 Spatial autocorrelation

The simplest method of estimating values at unknown positions from known sample values might be to average them, but this is sensible only if the values are independent of their location. Normally, however, a variable defined continuously over an



**Figure 15.1** Schematic histograms of a theoretical data population and theoretical sub-populations which could exist within the dataset

area does not change greatly from one place to another and so we can expect the unknown values to be related to those at nearby known points. This behaviour is described by Tobler's *first law of geography* (Tobler, 1970), which states that 'everything is related to everything else, but near things are more related than distant things'. The formal property that describes this is *spatial autocorrelation (SAC)*. Correlation represents the degree to which two variables or types of variables are related, while spatial autocorrelation represents the degree to which that correlation changes with distance. In the context of a raster image, this can be likened to making a copy of an image and overlaying the first one precisely; the two should be exactly correlated. If one image is then shifted by 1 pixel relative to the other and the correlation between them is examined again, they should still be very highly correlated. Continuing this process, shifting by 1 pixel and recalculating the correlation should lead to a point where the two images have been shifted so far that they are almost uncorrelated. If the collective results of this process are examined, a measure of SAC is produced. Understanding spatial autocorrelation is very useful since it can reveal and describe systematic patterns in the data which may not otherwise be obvious and may in turn reveal an underlying control on variation. Patterns which are truly random exhibit no spatial autocorrelation. SAC is important because it provides a measure of correlation in the dataset, and because it tests the assumption of randomness in the data. Here we recall the autocorrelation matrix introduced in Section 9.3.2, which characterizes the SAC in every direction.

Both positive and negative autocorrelation exist and are opposites of one another. *Positive spatial autocorrelation* occurs where near values or areas are alike and *negative spatial autocorrelation* where near values or areas are inversely correlated, while zero indicates no correlation between the two.

In general, two assumptions are made of *spatially autocorrelated errors*. The first is that the average error will be zero because the positive and negative fluctuations around the trend will cancel out one another. The second is that the precise locations of the errors are not significant, only their relative positions with respect to each other, a relationship known as *stationarity*.

### 15.2.3 Variograms

The variogram or more commonly the *semi-variogram* (half the variogram) is the main measure of similarity within a data population and is a principal tool of geostatistics. It is a statistical function that describes the decreasing correlation between pairs of sample values as the separation between them increases. Other tools such as the correlogram and *covariance* functions are also used but these are all very closely related to the variogram. A *covariance cloud* reveals the autocorrelation between pairs of data points, where each point in the cloud also represents a pair of points in the dataset. The cloud is generated by plotting the distance between the points against the squared difference between their values; points which are close together should also be close in value. As the distance between points increases, the likelihood of correlation between the point values decreases. The form of the cloud and the function fitted to it comprise the semi-variogram. The semi-variogram  $z(d)$  is then a function describing the degree of spatial dependence of a variable or process, and in general is defined by

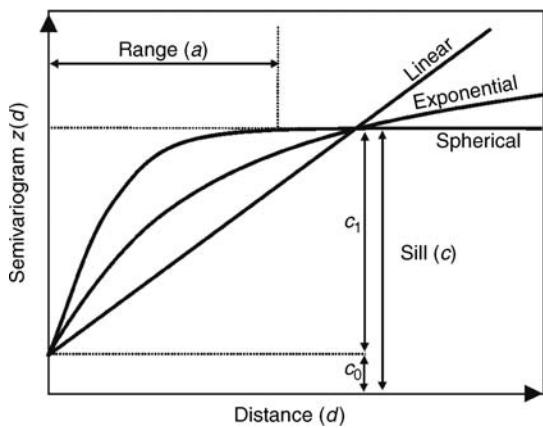
$$z(d) = \frac{1}{2n} \sum_{i=1}^n [z(x_i) - z(x_i + d)]^2 \quad (15.1)$$

where  $n$  is the number of sample pairs (observations of value  $z$  separated by distance  $d$ ) being evaluated, and  $x_i$  represents the positions of the points being compared. An idealized (theoretical) semi-variogram is a function defined by the relationship between the semi-variogram  $z(d)$  and distance ( $d$ ) between sample points; it is a plot of  $z(d)$  against  $d$ , as in Figure 15.2, and is used to describe data populations. There are several forms of theoretical semi-variogram and these can be defined as follows and illustrated schematically in Figure 15.2.

A *linear variogram*,  $z_l(d)$ , is a special and rather theoretical case, since it never reaches a sill, and is described by

$$z_l(d) = c_0 + pd \quad (15.2)$$

where  $c_0$  represents the nugget effect (which is random),  $p$  represents the gradient of the function, which is constant in the linear case, and  $d$  is the distance or *lag*. A *spherical variogram*,  $z_s(d)$ , can be used to fit data which reach a distinct sill. When  $z(d)$



**Figure 15.2** The form of several theoretical form of semi-variogram commonly used (linear, spherical and exponential, although there are several others), showing the relationship between the function, its sill ( $c$ , where  $c = c_0 + c_1$ ) and the distance ( $d$ ) at which the sill is reached ( $a$ ). Pairs of points plotted in the lower left-hand corner of the semi-variogram are close one another spatially; the opposite is true of points near the top-right corner of the semi-variogram

becomes constant and the sill roughly equals the calculated sample variance,

$$z_s(d) = c_0 + c_1 \left[ \frac{3d}{2a} - \frac{1}{2} \left( \frac{d}{a} \right)^3 \right] \quad \text{when } 0 \leq d \leq a$$

or  $z(d) = c_0 + c_1 \quad \text{when } d \geq a$       (15.3)

where  $a$  is the distance to the sill or *range*. If there is only a gradual approach to the sill, an *exponential variogram*,  $z_e(d)$  provides a good fit and can be described by

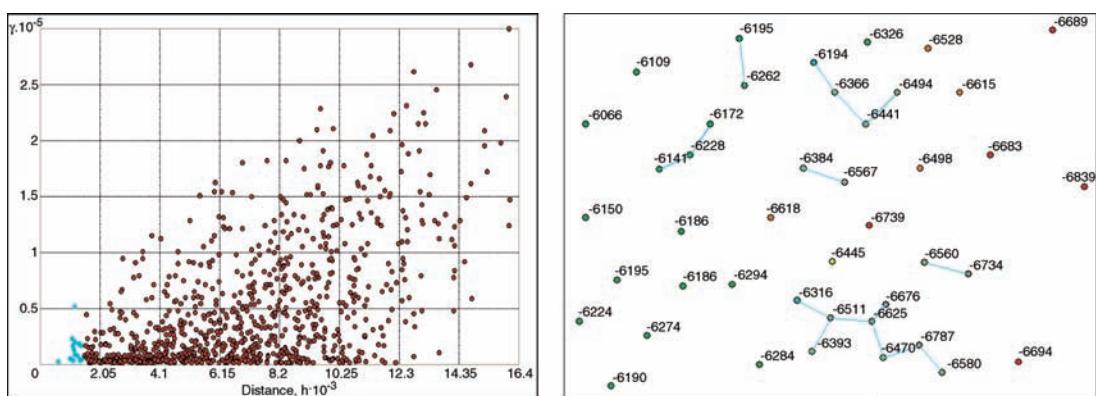
$$z_e(d) = c_0 + c_1(1 - e^{-d/a}). \quad (15.4)$$

The spherical and exponential variograms can be referred to as ‘transitive’ forms because the correlation varies with distance or lag ( $d$ ). Variogram forms which have no sill can be described as ‘non-transitive’ and the linear form can be used.

As illustrated in Figure 15.3, in a semi-variogram, pairs of points plotted in the lower left-hand corner are close one another spatially and have highly correlated values, and the opposite is true of points plotted near the upper right corner. The values of pairs of points plotted in the upper right of the semi-variogram (e.g. on the sill) can be considered to be uncorrelated. Like the histogram, the semi-variogram is useful for detecting *outliers* or pairs of points which have erroneous values, such as two closely adjacent points with wildly differing values.

#### 15.2.4 Underlying trends and natural barriers

The *trend* can be thought of as an underlying control on the overall pattern of the data values. If for instance the variable being predicted is elevation, the trend could be the regional slope gradient. In soil



**Figure 15.3** (a) Semi-variogram for a set of points (the  $y$  axis represents  $z(d)$  and the  $x$  axis the distance between points, as in Figure 15.2); and (b) the map of those points showing the relative positions (as cyan-coloured links) of the point pairs highlighted (also in cyan) in (a). The point values represent depth below sea level to a stratigraphic horizon and reveal a gently sloping surface; these values are used in the examples of interpolation types later on

or sediment geochemical data it could represent slope processes in moving debris (and elements) downslope under gravity. In airborne pollutants it could represent the prevailing wind direction. Such underlying trends may well affect the distribution of values and failure to consider them could produce misleading results. If the trend is not constant but variable, such as in the case of elevation data covering a sizeable area of terrain with, for example, a valley running across it, the function fitted to the data should allow for that. The assumption in this instance would then be that the mean is variable, and importantly that there is more than one statistically independent population present. Some methods of value estimation cannot make such allowances.

Physical, geographic barriers that exist in the landscape, such as cliffs or rivers, present a particular challenge when describing a surface numerically because the values on either side of the barrier may be drastically different. Elevation values change suddenly and radically near the edge of a cliff and the known values at the bottom of the cliff cannot be used accurately to estimate values at the top of the cliff. If natural barriers are known to exist in the data population then it will be advantageous to use a method of value extraction that can selectively use values on one side or the other. Many interpolators smooth over these differences by averaging values on either side of the barrier. The *inverse distance weighted (IDW)* method allows the inclusion of barriers to constrain the interpolation to one side.

## 15.3 Interpolation

Regardless of the quantity in question (rainfall intensity, pollution concentrations or elevation values), it is impossible or at best impractical to measure such phenomena at every conceivable location within an area. We can, however, obtain a sample of measurements from selected locations within that area, and from those samples make predictions about the values over the entire area. *Interpolation* is a process by which such predictions are made.

The process begins with a set of sample points containing numerical measurements recorded at

specific locations. Spatial autocorrelation is assumed so that an unknown value can be estimated from the neighbourhood of values. The aim is then to create a surface that models the sampled phenomenon so that the predicted values resemble the actual ones as closely as possible. Adjustments to the surface can be made by limiting the size of the sample used and controlling the influence that the neighbourhood of sample points has on the estimated values.

Interpolation can then be described as the process of estimating a value at a location  $(x, y)$  from irregularly spaced, assumed or measured values at other locations  $(x_1, \dots, x_n, y_1, \dots, y_n)$ , to produce a regularly or continuously sampled grid. It is possible to interpolate a surface from a very small number of sample points but more sample points will give a better result. Ideally, sample points should be well distributed throughout the area. If there are some rapidly changing phenomena, then denser sampling may be needed.

### 15.3.1 Selecting sample size

This is an important step as it controls the neighbourhood statistics from which the interpolated values will be estimated. Most interpolation methods allow you to control the number of sample points used, in some way or other. For example, if you limit your sample by number, to five points for example, for every location the interpolated value will be estimated from the five nearest points. The distance between each sample point varies according to the distribution and density of the points, as we have said, and this distance is important. Using many points will slow the process down but will mean that the distances between the points are smaller, so variation between them will be lower and the result should be more accurate. Using fewer points will make the process faster, and sufficient points are likely to be found, but the prediction may not properly represent the statistics of the neighbourhood. The sample size can also be controlled by use of a *search radius* or by defining the minimum number of points to be used.

Two common approaches to sample selection are the *fixed distance method* and the *nearest K*

*neighbours method* (where  $K$  is a specified number). A *fixed search radius* will use only the samples contained within the specified radial distance of the unknown value, regardless of how large or small that number that might be. The  $K$  nearest neighbour method uses a *variable search radius*, which expands until the  $K$  neighbouring points are found. The fixed distance technique, shown in Figure 15.4a, using a distance equal to the radius of the circles shown, would interpolate the value at point B using four neighbouring samples but would find only one sample to interpolate the value at point A. If, instead, a variable radius were used (Figure 15.4b), then the search around point A would have to expand considerably before four neighbours are found.

The fixed distance approach may fail to find any sample points and the interpolator will fail to estimate a value within an area of low-density sampling. This is useful only in that it will reveal areas where there is insufficient sampling, but the resultant ‘holes’ in the interpolated surface are rather undesirable. The  $K$  nearest neighbour approach, in contrast, will always find sample points but they may be so far from one another (as in Figure 15.4b) as to be unrelated and so the predicted result may be misleading.

Clearly, the choice of method will depend on the data, how they were collected, the desired characteristics of the output grid, and the nature of the decisions or analyses that will be performed with the resulting grid. Once a set of neighbours is found, the interpolator must combine their values to produce the estimate.

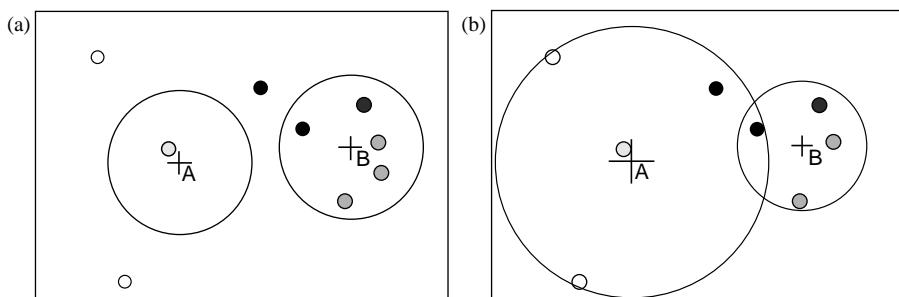
### 15.3.2 Interpolation methods

There are two broad classes of interpolator, *deterministic* and *stochastic*. A *deterministic process* is one where, at any specific known instant, there is only one possible outcome. In general terms, a *stochastic process* on the other hand exhibits probabilistic behaviour, i.e. it can be considered the opposite of deterministic, so that for one known condition there are many possible outcomes, some of which will be more likely than others. Deterministic interpolators create surfaces based either on the degree of similarity between sample values (as in the IDW method) or on the degree of smoothing (as with *radial basis functions (RBFs)*). Stochastic interpolators are based on the degree of autocorrelation between every possible combination of points in the input dataset. It is generally considered that, in situations where data are plentiful, stochastic interpolation methods are superior. A summary of different interpolation methods is shown in Table 15.1.

### 15.3.3 Deterministic interpolators

The majority of deterministic interpolators are polynomial in form, and of varying degrees of complexity. The general form of a polynomial function is

$$f(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x^1 + a_0 x^0 \quad (15.5)$$



**Figure 15.4** Neighbourhood search methods using (a) a fixed distance and (b) a variable distance to find the  $K$  nearest neighbours, to estimate two points located at A and B

**Table 15.1** A selection of interpolators compared

Class	Type	Uses SAC	No. of variables	Honours the data	Surface Type	Pros	Cons	Assumptions
Deterministic	Polynomial (global)	No	1	Yes	Prediction only	Simple	Too simple	None
	RBF					Poor in high relief data	Poor in high relief data	
	IDW	No	2+	Depends	Simple & barriers	'Bulls eyes'	Poor in sparse data	Stationarity
	Kriging					Very flexible & allows for trends		
Stochastic	Co-kriging	Yes	2+	Depends	Prediction, probability, quantile & standard error	Poor in sparse data	Poor in sparse data	

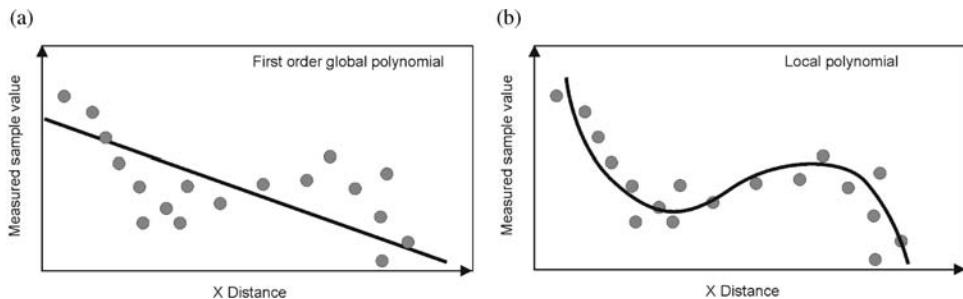
SAC – Spatial autocorrelation, RBF – Radial Basis Functions, IDW – Inverse Distance Weighted

where  $x$  is the input value, the number of terms is variable and each term consists of two factors (a real number coefficient  $a$  and a non-negative integer power  $n$ ). The degree or order of a polynomial function is given by the highest value of  $n$ .

*Global polynomial* interpolators use a polynomial function to construct a very simple surface from all the sampled point values, i.e. no neighbourhood is specified, and so smoothes over all local variations. First-, second-, third- or fourth-order polynomials, and so on, can be used to represent surfaces of increasing complexity. A smooth plane is created with a first-order, a surface with one bend or fold is made by a second-order, and one with two bends or folds by a third-order polynomial. Since the surface is relatively rigid it will not honour the data, i.e. it will not necessarily pass through all the data point values. Global polynomial interpolators are often referred to as ‘inexact’ interpolators for this reason.

For a set of sample points representing surface elevations, shown in Figure 15.3b, the result of interpolating using a first-order global polynomial is shown in Figure 15.5a.

If a specific neighbourhood is then selected, the result becomes equivalent to a *local polynomial interpolator*. Local polynomial interpolation creates a surface using functions unique to a sample neighbourhood. By controlling the number of points, the shape of the neighbourhood and the location of the points within the neighbourhood (i.e. the sector configuration), even more control is enabled. In this way the interpolation can be made to behave in a more (or less) local manner. The process is a little like convolution filtering, in that a function is fitted to the values in a neighbourhood to derive an estimated value for that unknown location. The interpolator then shifts to the next unknown location and the process is repeated until a grid of estimated values is built up. Figure 15.5b shows the result of local polynomial interpolation for the same group of points. As with the global interpolator, selection of first-, second-, third-order, and so on, polynomial functions allows more complexity to be allowed for in the predicted surface except that these are fitted within the local neighbourhood. Hence if the neighbourhood size is increased to the point where it includes all the data points, the result will



**Figure 15.5** (a) A first-order global polynomial surface profile. The planar surface (black line) does not pass through the sampled points and reflects only the gross scale pattern of the data. (b) A local polynomial surface profile. The surface is no longer planar but has flexure; it still does not pass through all the data points

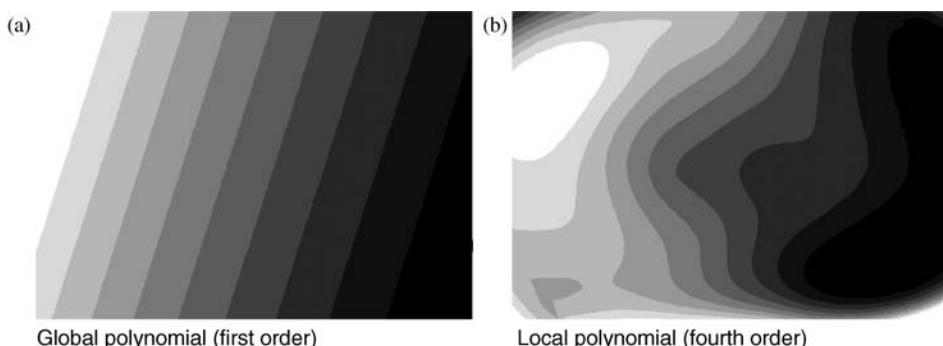
be equivalent to a global polynomial interpolator (Figure 15.6).

By adding more orders to the polynomial, it can be made to fit almost any data distribution, but if the data are not that complex, then why bother going to such effort when a simpler one would be quicker and more appropriate? If you have to work so hard as to make the function fit the data, that extra effort may not really provide much additional information, and perhaps this is telling you something about the data you have overlooked, such as the presence of an underlying trend and/or more than one population. Generally speaking, and for these reasons, the first and second orders of the polynomial are considered to be the most indicative and significant in fitting to the data and they are in fact estimates of the first- and second-order trends of the data.

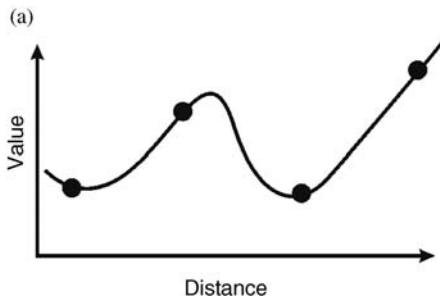
RBFs and *splines* generate surfaces using a piecewise function which can be thought of as a flexible

membrane rather than a rigid plane, stretched between the sample points, of which the total curvature is kept to a minimum but is variable (Figure 15.7). Variable weights are used to control the flexibility and curvature of the interpolated surface. The surface passes through every sampled point, and so spline functions can be described as being ‘exact’ interpolators. The stretching effect is useful as it allows predicted values to be estimated above the maximum or below the minimum sampled values, so that highs and lows can be predicted when they are known to exist but are not represented in the sample data.

When sample points are very close together and have extreme differences in value, spline interpolation is less effective because it involves slope calculations and honours the data. High-frequency changes in value, as caused by a cliff face, faults or other naturally occurring barriers for example,



**Figure 15.6** Surfaces constructed from the sample points shown in Figure 15.3b, using (a) global polynomial and (b) local polynomial functions



**Figure 15.7** Profile view of a theoretical surface constructed with a typical spline interpolator

are not represented well by a smooth-curving surface. In such cases, the IDW method may be more effective.

Several types of spline interpolator can be found in most GIS suites: tension, thin-plate (minimum curvature), regularized, multi-quadric and inverse multi-quadric splines. A *tension spline* is flatter and more rigid than a regularized spline of the same sample points – it forces the estimated values to stay closer to the sampled values. The behaviour of a *regularized spline* is more flexible and elastic in

character (it has greater curvature). Interpolated surfaces generated from the points in Figure 15.3 are shown in Figure 15.8.

The spline interpolation process can also be weighted. The simplest form of RBF is a weighted linear inverse distance function as follows:

$$z_p = \sum_{i=1}^n \lambda_i z_i \quad (15.6)$$

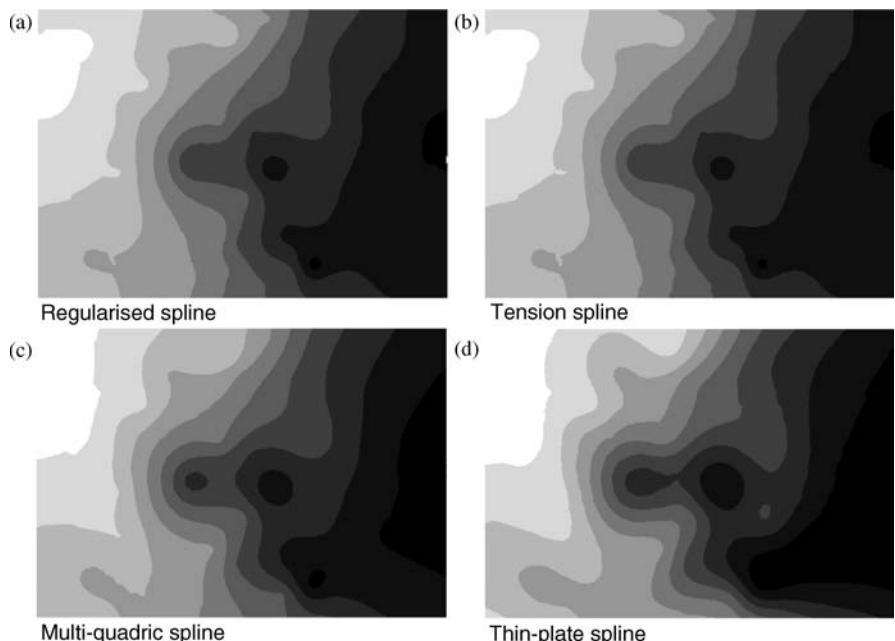
where  $z_p$  is the estimated value of the interpolated surface at a point  $p$ , and  $\lambda_i$  are the data weights.

The form of a tension spline can be expressed by

$$\varphi(r) = \ln(cr/2) + l_0(cr) + \gamma \quad (15.7)$$

where  $\varphi(r)$  is the RBF used,  $r$  is the distance between the point and the sample,  $c$  is a smoothing parameter,  $l_0()$  a modified Bessel function and  $\gamma$  is Euler's constant ( $\gamma = 0.577$ ). The modified Bessel function is given by

$$l_0(cr) = \sum_{i=0}^{\infty} \frac{(-1)^i (cr/2)^{2i}}{(i!)^2}. \quad (15.8)$$



**Figure 15.8** Surfaces constructed from the same set of sample points, shown in Figure 15.3b, using (a) regularized; (b) tension; (c) multi-quadric; and (d) thin-plate splines. The differences between these are subtle, with surfaces in (c) and (d) being noticeably smoother than those in (a) and (b)

The general form of a regularized spline can be described by

$$\varphi(r) = \ln(cr/2)^2 + E_1(cr)^2 + \gamma \quad (15.9)$$

where  $E_1()$  is an exponential integral function given by

$$E_1(x) = \int_1^{\infty} \frac{e^{-tx}}{t} dt. \quad (15.10)$$

A *multi-quadratic spline* is defined as

$$\varphi(r) = \sqrt{r^2 + c^2} \quad (15.11)$$

and a *thin-plate spline* as

$$\varphi(r) = c^2 r^2 \ln(cr). \quad (15.12)$$

In the regularized type the predicted surface becomes increasingly smooth as the weight value increases. With the tension type, increasing the weight produces a more rigid surface, eventually approaching a linear interpolation between sample point values. Splines cannot assess prediction error, cannot allow for SAC and do not involve any assumptions about the stationarity of the data.

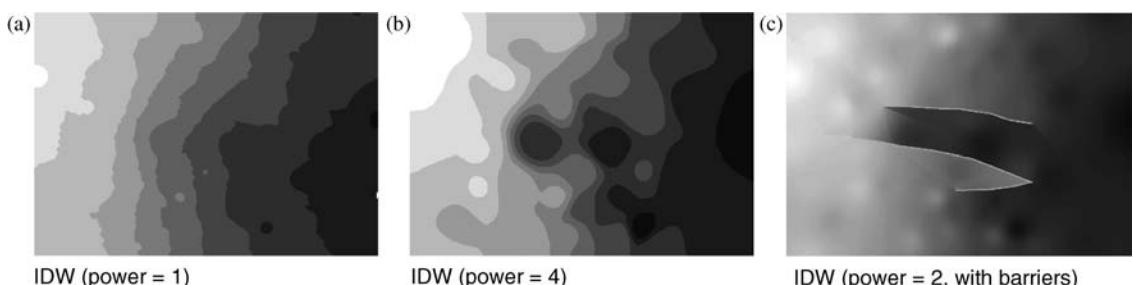
Minimum curvature spline interpolation has been used for the ‘gridding’ of multi-element geochemical sample points in Greenland, within a multi-criteria evaluation of mineral prospecting in Section 22.1.

### 15.3.3.1 Inverse distance weighted average

The IDW average is a localized interpolator that predicts values through averaging, as its name suggests, but allows variable *weighting* of the averages according to the distances between the points, using a power setting. The weights are exponents of distance and are largest at zero distance from a location and decrease as the distance increases. For a position  $x$ , and for  $i$  to  $n$  data points with  $z$  known values, the unknown weighted average ( $z(x)$ ) is derived as

$$z(x) = \frac{\sum_i w_i z_i}{\sum_i w_i}. \quad (15.13)$$

Reducing the weight produces a more averaged prediction because distant sample points become more and more influential until all sample points have equal influence (Figure 15.9a). Increased weight or power means that the predicted values become more localized and less averaged, but the influence of the sample point decreases more rapidly with distance (Figure 15.9b). The weights are commonly derived as the inverse square of distance, so that the weight of a point drops by a factor of 4 as the distance to a point increases by a factor of 2. For the sample points shown in Figure 15.3, surfaces interpolated by IDW, with low and higher power setting, are shown in Figure 15.9.



**Figure 15.9** Surfaces constructed from the same set of sample points, shown in Figure 15.3b, using (a) IDW with a low weight setting; (b) IDW with a high weight setting; and (c) with break-lines. The interpolated grids in (a) and (b) were produced using ArcGIS’s Geostatistical Analyst and in (c) using ArcGIS’s Spatial Analyst; the ‘bullseye’ pattern is visible in all three grids

Since the IDW is an averaging technique, it cannot make estimates above the maximum or below the minimum sample values, and as a result the predicted surface will not pass through the sample points, so it can be referred to as an ‘inexact’ interpolator. In a surface representing elevation, for example, this has the effect of flattening peaks and valleys unless their high and low points are actual sample points (see Figure 15.10).

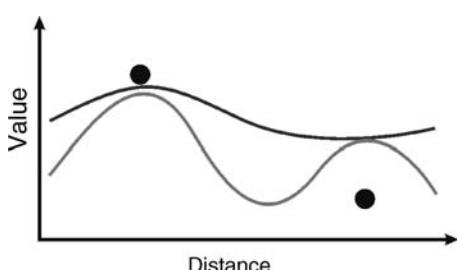
One advantage of this method is that barriers can be incorporated to restrict the predictions, geographically, if structures are known to exist which affect the shape of the surface. Sample points on one side of a barrier are excluded from the interpolation even if they are near to the prediction location (see Figure 15.11). In this way the IDW is prevented from averaging across significant structures.

The IDW is most effective with densely and evenly spaced sample points. It cannot account for any directional trends in the data, and so the interpolated surface will average across any trend rather than preserve it. It is perhaps useful as a ‘first attempt’ when little is known about any complexities which may exist. It does not involve the assessment of prediction errors or allow for SAC, and the weighting tends to produce ‘bullseyes’ around sample point locations.

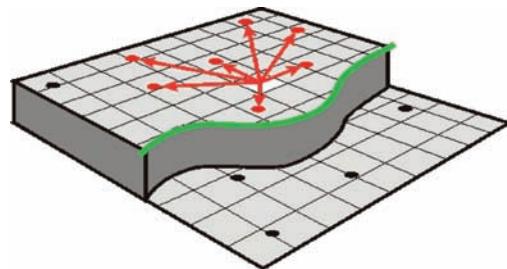
#### 15.3.4 Stochastic Interpolators

##### 15.3.4.1 Kriging

The kriging method was first developed by Matheron (based on the work of Krige, 1951). It



**Figure 15.10** Profile view of a theoretical surface constructed with an IDW interpolator: the black line represents the surface generated with a lower weight setting; the grey line represents that produced with a higher weight setting (compare with Figure 15.9a and b)



**Figure 15.11** Schematic representation of sample points lying across an abrupt change in values, such as might be caused by a cliff within elevation data. The IDW interpolator allows incorporation of vector break-lines (green line) to constrain the interpolation process, so that only points on one side (red dots) are used to estimate the output value (white pixel)

is a method of estimation based on the trend and variability from that trend. *Variability*, in this context, refers to random errors about the trend or mean. In this context, ‘error’ does not imply a mistake but a fluctuation from the trend, and ‘random’ implies that the fluctuation (error) about the trend is unknown and is not systematic; the fluctuation could be positive or negative. Kriging may be considered exact (or smoothed) or inexact. Kriging incorporates the principles of probability and prediction, and like the IDW, is a weighted average technique except that a surface produced by kriging may exceed the value range of the sample points while still not actually passing through them. Various statistical models can be chosen to produce four map outputs (or surfaces) from the kriging process. These include the interpolated surface (the prediction), the standard prediction errors (variance), probability (that the prediction exceeds a threshold) and quantile (for any given probability).

Simplistically, all forms of kriging are based on the following relationship:

$$z_{xy} = \mu_{xy} + \varepsilon_{xy} \quad (15.14)$$

where  $z_{xy}$  is the predicted surface variable (at location  $xy$ ),  $\mu_{xy}$  is the deterministic mean or trend of the data and  $\varepsilon_{xy}$  is the spatially autocorrelated error associated with the prediction.

The general form of kriging can be defined as

$$\hat{f}(z_{xy}^o) = \sum_{i=1}^n w_i f(z_{xy}^i) \quad (15.15)$$

where the function determines the output prediction value of a location so that  $\hat{f}(z_{xy}^o)$  is the predicted output value and is a weighted linear combination of the input values (ranging from  $i$  to  $n$ ), and  $w_i$  refers to the weight for the  $i$ th input value.

So the predicted surface value at any position is a function of the trend and the deviation from that trend. The differences between the different forms of kriging can be explained in reference to this relationship. *Ordinary kriging* assumes an unknown but constant mean, i.e.  $\mu_{xy} = \mu$  at all locations, so there is no underlying trend to the data, and that the sample values are random (spatially autocorrelated) errors about the unknown mean (Figure 15.12a). In situations where there is a trend and the mean is no longer constant, the trend is represented as a linear or nonlinear regression; this is the basis of *universal kriging*, which assumes a varying but still unknown mean and that the sample values are random (spatially autocorrelated) errors about the mean (Figure 15.12b). In contrast, *simple kriging* assumes that the mean is known in advance and that it may be constant or variable. *Indicator kriging* involves the use of other transformations,  $f(z_{xy})$ , applied to the predicted value rather than the sample value, such that the predicted values are signed (0 or 1) representing the probability that the surface value will exceed or fall below a specified threshold: 1 if the value is above the threshold, 0 if below. This may be useful if predicting values on which rigid decisions

will be made, such as whether chemical substances are in high enough concentrations to warrant an area being classified as contaminated or not. It can be thought of as a combination of kriging with reclassification, and forms an area of overlap with multi-criteria evaluation. *Disjunctive kriging* forms a development of this approach in which a series of possible transformations is searched to predict the function of  $f(z_{xy})$ .

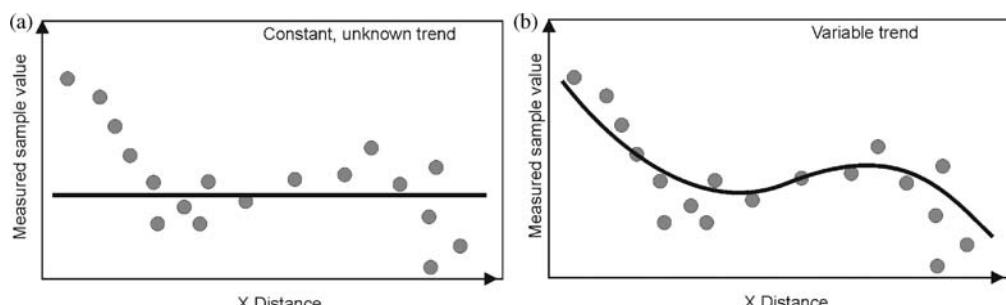
Kriging assumes stationarity in the data and in some methods that the data are normally distributed.

#### 15.3.4.2 Co-kriging

Where kriging involves interpolation of a single variable, *co-kriging* involves the simultaneous interpolation of more than one variable. As a result, co-kriging allows the derivation of cross-correlation as well as SAC, and is given as a minor modification of (15.13):

$$z_{xy}^i = \mu_{xy}^i + \varepsilon_{ixy}^i. \quad (15.16)$$

In this way, different trends and SAC can be considered for each of the  $i$  variables. This may be useful if you do not have equal number of sample points for all variables and need to share values; the prediction can be made from the values of both variables and from the correlation between them. For instance, if you have multi-element geochemical data for samples collected by different ground sampling strategies (rock *in situ* samples, rock transport, sediment, etc.) but not many samples for any single collection method, it may be useful to interpolate the concentrations from all available sample types for a particular element. Care must



**Figure 15.12** Ordinary and universal kriging (illustrating constant mean and varying mean). (a) Ordinary kriging with a constant mean (no trend) when the mean value is not known in advance; and (b) universal kriging in which there is a trend in the data, but the terms of its function are not known

be taken, however, to ensure that the combination of sample points is conceptually meaningful (i.e., that the two sample types being combined represent the same or comparable phenomena). Alternatively, if particular associations of elements are indicators of some phenomena of interest, co-kriging using those elements and evaluating the relationships between them (and their error patterns) may be very revealing. Here again there is an overlap with multi-criteria evaluation and decision making.

All the forms of kriging product are also produced by co-kriging, creating ordinary co-kriging, universal co-kriging, simple co-kriging, indicator co-kriging and disjunctive co-kriging.

## 15.4 Summary

The significant aspects to take away from this chapter include the importance of knowing your data from the start, understanding their make-up and provenance, and so to make the best choice of interpolator. Understanding how the data were collected, recorded and measured, point to what they represent, or, importantly, what they do not represent, ‘on the ground’. This is essential to the understanding of any statistics derived from the data. Realizing the existence of populations within the data is important when interpolating, since treating the data as a single population when in fact there are several could produce meaningless results and could mean the underestimation of calculations or forecasted quantities. Similarly, determining how points are selected for the interpolation process can have significant effects on the validity of the result.

In terms of control over the interpolation process, RBFs or splines can be considered more flexible than the IDW and less flexible than kriging, but it is the distribution and quantity of your data which should dictate the kind of interpolation you use rather than the convenience of the tools. In general,

when data are plentiful, geostatistical methods give better results and, unlike the simpler methods, do not treat noise as part of the data.

This chapter also touches on the conversion of vector point data into a raster representation; that is, estimating values where none exist. There are also instances where the concepts here have elements in common with decision making, multi-criteria evaluation and spatial analysis.

## Questions

- 15.1 What methods of sample selection are available? What potential effects of the methods used?
- 15.2 What are the advantages and potential dangers of using interpolation?
- 15.3 Under what circumstances might you decide it was not appropriate to use interpolation?
- 15.4 When are spline interpolation methods most and least useful?
- 15.5 Why would you need to include barriers in interpolation methods?
- 15.6 What are the differences between kriging and other weighted methods of interpolation? When is kriging not likely to give better results than any other types?
- 15.7 Why is it important, in kriging, to have a thorough understanding of the variance (and variogram) of the data?
- 15.8 Why is it worth using interpolated data (think about explaining this to a decision maker)?
- 15.9 Consider some applications and decide which interpolation methods would be most appropriate – such as assessing the probability of geochemical contamination from regularly spaced soil samples, or estimating the production of a mineral commodity, using geochemical data derived from various samples of rock and sediment collected from several levels within a mine.



# 16

## Representing and Exploiting Surfaces

### 16.1 Introduction

A *surface* models a phenomenon that varies continuously across an area, such as elevation. Since the phenomenon could represent precipitation, temperature or magnetic susceptibility, or any variable, a more general term could be a *statistical surface* because the surface describes the statistical representation of the magnitude of that variable. Surfaces provide the ‘height’ information, or *z* values, necessary both for spatial analysis and for 3D visualization, in the form of either raster DN values or the nodes of a Triangulated Irregular Network (TIN).

This chapter concentrates on the use of raster data since their structure lends itself to terrain analysis, allowing the description and quantification of terrain morphology and the extraction of surface parameters in a more uniform, regular manner than from a vector surface. This chapter also deals with the visualization of information which has an implied 3D quality within a 2D environment, e.g., on a map, the visualization of information within a simulated 3D environment, and the exploitation of surface data to derive parameters that quantify the 3D environment and are useful within the broader scope of spatial analysis. When talking about raster data visualization, image processing techniques are embedded by default.

### 16.2 Sources and uses of surface data

Methods of surface description using raster and vector models have been mentioned earlier in Chapter 11. Primary sources of surface data include point surveys, *photogrammetry* using stereo imagery or aerial photography, *interferometry* from radar imagery, and *altimetry*. Surfaces can also be produced by the digital capture of contours from analogue maps (and secondary conversion to a surface) and by interpolation from survey points. We will not dwell on the use of contours (as these are familiar concepts which have been dealt with in many other texts) and interpolation from point data has already been dealt with in Chapter 15.

#### 16.2.1 Digital Elevation Models

‘Digital Elevation Model’ or DEM is a term used to describe a representation of a continuously sampled surface representing ground surface height above a datum. A DEM generally represents the uppermost level of a surface feature, including vegetation canopies and buildings; it does not necessarily represent the ground surface level of the Earth. If this is required, the DEM must be modified

to remove any such building and tree canopy heights. And when this is achieved, the product may be described as a ‘bare Earth model’, Digital Terrain Model (DTM) or Digital Surface Model (DSM). The term ‘bare Earth’ can refer to either the DEM or the extracted contours, from which the effects of objects such as buildings and tree canopies have been removed, leaving only ground surface elevation values. The production of the DTM from the DEM requires either post-processing to correct to a bare Earth model, or calculation from the raw acquired elevation data, as in the case of laser altimetry where the collected data represent complex information containing the uppermost surface and ground-level elevations (and any other objects in between). For the sake of simplicity in this chapter we will stick to the acronym ‘DEM’ as a generic term in reference to all forms of digital elevation data.

A DEM is therefore a digital data file consisting of terrain elevations for ground positions at regularly spaced horizontal intervals. DEMs may be used aesthetically or analytically; they can be used in combination with digital images and vectors to create visually pleasing and dramatic graphics or for the calculation of various surface parameters such as terrain slope, aspect or profiles.

One potentially misleading issue relates to the term ‘continuous’ which is frequently used to describe raster data. It is more correct to describe the model as being ‘continuously sampled’ at discrete intervals rather than truly continuous in nature. Its ability to represent a surface depends on its spatial resolution and the complexity of the ground surface being represented. It is widely accepted that all natural surfaces are fractal in nature, so that at any particular scale, there will always be more detail than we can observe. We must therefore accept that the resolution of the data implies the level of detail that we can work with. The most important factors to be considered prior to the use of DEMs therefore include the planimetric and altimetric accuracies of the source data, the quality and quantity of both the source data and ground control data, the level of terrain complexity, the output spatial resolution as well as the algorithm used to generate the DEM.

The many processes of DEM generation are the subject of a great breadth of research and are covered in great detail in other texts. The technical

details of such procedures are not within the scope of this book and so we give only a brief overview here. DEMs can conveniently be subdivided according to source data, method of generation and/or mapping scale of application; here we have chosen to group them according to the generation method. The primary generation of surface information can be achieved through photogrammetry (from stereo images), altimetry and laser altimetry (LiDAR) and SAR interferometry (InSAR).

#### 16.2.1.1 Photogrammetry

Photogrammetry is the process of obtaining reliable 3D measurements of physical objects and the environment from measurements made from two or more photographs or images (Wolf and Dewitt, 2000). The photographs or images must have been acquired from different positions with sufficient overlaps, i.e. stereoscopically. Two forms of photogrammetry can be identified: metric and interpretative. The former refers to the quantitative measurement and analysis of objects for the purpose of calculating dimensions, including elevation and volume. The latter refers to the more qualitative interpretation and identification of objects and structures through analogue stereoscopy, with the aim of better understanding their relationships with their surroundings.

Elevation data can be derived photogrammetrically from a number of readily available data sources (airborne or spaceborne) such as stereo areal photography, ASTER, Ikonos, SPOT, EROS, Orbview, Topsat, WorldView and, most recently, the ALOS and GeoEye1 satellite sensors. The height accuracy of models generated in this way depends mainly on the *base to height ratio (B/H)* and the accuracy of the parallax approximations. The sensor specifications of these instruments are given in Appendix A.

ASTER (Advanced Spaceborne Thermal Emission and Reflection Radiometer) provides along-track stereo capability in addition to its spectral and spatial capabilities. It consists of three subsystems in the visible and near-infrared (VNIR), short-wave infrared (SWIR) and thermal infrared (TIR) regions, with 15 m, 30 m and 90 m resolutions respectively. The VNIR subsystem contains two independent telescope arrays: one at nadir (i.e. vertical looking) and a second, backward-looking telescope. These two arrays provide the along-track stereo image geometry, with a base to height

ratio of 0.6, and an intersection angle of 27.7°. ASTER data are relatively inexpensive (per km<sup>2</sup>) and a full scene covers an area of 60 × 60 km. ASTER DEMs can be produced at 15, 30 or 60 m resolutions and are suitable for mapping at a scale of about 1 : 50 000 (Welch *et al.*, 1998). ASTER is, however, and rather unfortunately, approaching the end of its working life; we await news of its replacement.

*SPOT* (Satellite Pour l'Observation de la Terre): Stereo pairs from SPOT 2–4 are acquired from cross-track imaging with sensors viewing the same area off nadir (see Appendix A). The data can be used to produce 10 m DEMs, while data from SPOT 5 enables 5 m DEMs to be generated. The SPOT 5 HRS sensor routinely acquires large volumes of stereo data, so that DEM products can be obtained on a 'per unit area' basis. Relative DEMs (i.e. produced using no ground control data) of small areas of interest can be produced, which have planimetric accuracies of ±30 m. The accuracy of absolute DEMs varies according to the quality of the ground control data but may be between 15 m and 30 m (planimetric) and 10–20 m (altimetric).

*Ikonos*: Cross-track stereo Ikonos images (1 m panchromatic and/or 4 m multi-spectral) are acquired at off-nadir angles of up to 60° in any direction. The sensor has 360° pointing capability, can acquire stereo images between latitudes of ±82° and can achieve base to height ratios of 0.6 or more (similar to those typical of aerial photography). Planimetric and altimetric accuracies range between 1 m and 2 m (depending on the accuracy of the ground control).

The *GeoEye-1* VHR sensor launched in September 2008 is similar to Ikonos but with some minor wavelength modifications, slightly wider swath and off-nadir spatial resolutions of 0.41 m (in panchromatic mono and stereo modes) and 1.64 m (multi-spectral mode). It is capable of viewing and imaging in any direction; with off-nadir look angles of 10°, 28° and 35° it can acquire cross-track stereo imagery with a horizontal error of 2 m (circular error (CE), i.e. error in all directions) or 3 m (linear error) with no ground control. It also provides revisit times of between 2.1 and 8.3 days, depending on the viewing angle.

The *WorldView-1* VHR sensor is capable of along-track stereo data acquisition, collects pan-

chromatic imagery at 0.5 m spatial resolution and is capable of acquiring imagery with accuracies of between 4 m and 5.5 m CE at nadir.

The *ALOS* (Advanced Land Observing Satellite) sensor assembly carries the Panchromatic Remote-sensing Instrument for Stereo Mapping (PRISM) which is capable of acquiring along-track stereo imagery using three optical devices that point forward, nadir and aft, and at a spatial resolution of a few metres (2.5 m at nadir). The advantage of three-way stereo is to minimize occlusion.

*Stereo aerial photography*: The resolution of DEMs derived from aerial photography can vary greatly according to the flight height, camera quality and imaging configuration. Aerial photographs are now collected using high-resolution digital cameras, and with onboard Global Positioning System (GPS) devices to georeference the acquired data and with inertial navigational units (to record and subsequently correct for the roll, pitch and yaw of the platform). Unmanned airborne vehicles (UAVs) are also now increasingly being used to map large areas at very high resolutions, and in stereo.

#### 16.2.1.2 Laser altimetry

Surfaces can be generated from laser altimetry or *light detection and ranging* (LiDAR) data, which may be acquired from satellite or airborne platforms. Airborne LiDAR has somewhat revolutionized the acquisition of high-accuracy DEM data for large-scale mapping applications.

A LiDAR system transmits pulses of light which reflect off the terrain and other ground objects. The receipt of laser pulses is continuous and so the first and last returned pulses can be extracted to differentiate between canopy elevations and true ground or bare Earth elevations. The return travel time of a laser beam is measured from the source instrument to the target and back; the distance is then computed (using the known speed of light) to give the height of the surveyed ground position. An airborne LiDAR system typically consists of a laser scanning instrument, a GPS and an inertial navigational unit. Airborne LiDAR derives height elevations with accuracies of between 10–15 cm (altimetric) and 15–30 cm (planimetric).

With the same ranging principle, airborne radar altimeters and barometric altimeters are also used to

map terrain. The radar altimeter height can then be subtracted from the barometric altimeter height to give surface elevation with respect to sea level.

Some LiDAR datasets are available as off-the-shelf products, from satellite sources such as ACE, a 1 km (30 arc second) DEM, globally available from De Montfort University (see online resources). These are generally low-resolution products aimed at small-scale regional applications. Airborne LiDAR surveys are normally bespoke and relatively expensive to commission and are therefore not likely to be freely available in the foreseeable future.

#### 16.2.1.3 SAR interferometry

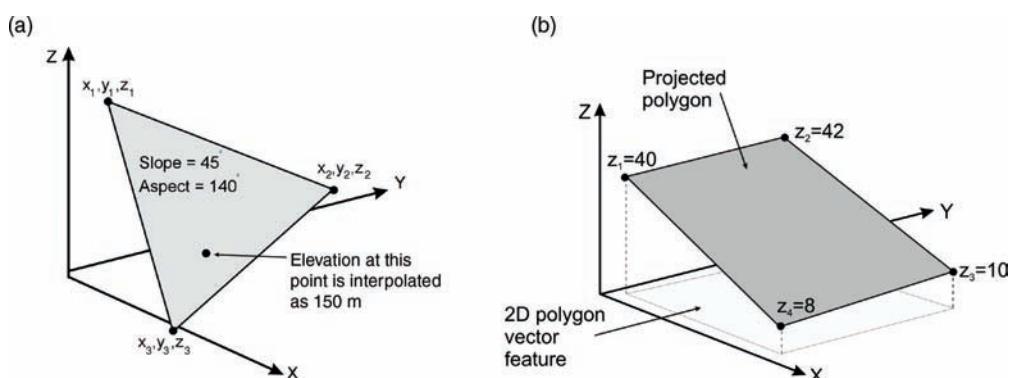
The theoretical basis for interferometric derivation of elevation from SAR data has already been described in some detail in Chapter 10, so we have no need to dwell on this here. InSAR DEMs can be generated from SAR imagery acquired by the ERS, ENVISAT, RADARSAT and ALOS satellites. The unique advantage of the InSAR DEM is its all-weather capability. It can penetrate clouds under conditions where all optical-sensor-based technology cannot operate.

One very well-known and widely used such InSAR DEM dataset was produced during the Shuttle Radar Topographic Mission (SRTM). SRTM was flown during the year 2000, onboard the Space Shuttle *Endeavour*, when topographic data of roughly 80% of the Earth's land surface (between

latitudes 60°N and 60°S) were generated and then gridded at 90 m (a higher resolution product is available for the United States). The stated absolute planimetric and altimetric accuracies of the 90 m DEMs are 20 m and 16 m respectively. The SRTM product is ideal for regional mapping, typically at scales of between 100 000 and 150 000.

#### 16.2.2 Vector surfaces and objects

As described earlier in Chapter 12, TINs represent surfaces using a set of non-overlapping triangles that border one another and vary in size and form. TINs are created from input points with  $x$ ,  $y$  coordinates and  $z$  values. The input points become the triangle vertices (nodes) and the vertices are connected by lines that form the triangle boundaries (edges), as illustrated in Figure 16.1. The final product is a continuous surface of triangles, made of nodes and edges. This allows a TIN to preserve the precision of the input data while simultaneously modelling the values between known points. Any precisely known locations such as spot heights at mountain peaks, or road junctions, can be described and added as input points, to become new TIN nodes. TINs are typically used for high-precision modelling, in small areas, such as within engineering applications, where they are favoured because they honour the data.



**Figure 16.1** (a) Individual triangular face of a TIN, defined by the 3D coordinate positions of the three points. The slope and aspect of each face are constant for each triangular face, and are calculated when the TIN is generated. The elevation of any position is then interpolated from its position with respect to the points and edges of the triangular face. (b) The 3D vector feature formed by projection of vertices using stored attribute ( $z$ ) values; these values could represent either the base-level elevation of the object or the object's height above ground level or any other quantity described in the attribute table

Once the TIN is constructed, the elevation of any location is interpolated from the  $x$ ,  $y$ ,  $z$  coordinates of the bounding triangle's vertices. The slope and aspect are constant for each triangular face and calculated during TIN construction (Figure 16.1).

Since elevation, slope and aspect are built into the structure of a TIN, they can very easily be displayed simply by symbolizing the TIN faces using these attributes. Four different edge types are stored within TINs; these may be referred to as *hard edges* (representing sharp breaks in slope or elevation), *soft edges* (representing gentle changes in slope), *outside edges* (generated automatically when the TIN is created, to constrain or close the outer triangles), and *regular edges* (all remaining edges other than the hard, soft or outside). A TIN surface can be displayed merely as edges, nodes or faces, or as combinations of these features, to enable the surface to be interpreted and any potential errors identified.

#### 16.2.2.1 3D vector features

Capitalizing on the TIN method of height representation, ordinary 2D vector features (points, lines or polygons) can be displayed in a 3D environment, on, above or beneath a surface. A 3D display requires height information but vector features may or may not have such information in their attribute tables. Truly 3D vector features have  $z$  values stored in their attribute tables but 2D vector features can also be visualized in 3D space by exploiting the height information of other data layers. The elevations are then calculated for the  $x$  and  $y$  positions of the vector feature's vertices and used to project them vertically. These heights, exploited from other layers or otherwise, are generally assumed to define the base level or ground height of the object.

A feature may of course store further attributes which represent the height of the object above ground level, such as building height, or any other quantity. A feature which contains, in addition to its  $x$ ,  $y$  coordinates, one or more  $z$  values as part of its geometry is referred to as a *3D vector feature*. At their simplest, 3D points have one  $z$  value; 3D lines and polygons have a  $z$  value for each vertex defining the object. Buildings and other, far more complex structures can then be modelled by extending this concept.

VRML (Virtual Reality Modelling Language) was developed as a standard file format for storing

and displaying 3D vector graphics. This has been superseded by X3D, which is now an ISO standard; both formats use standard XML syntax. The development of 3D vector models and vector topology is a very active area of research and one which is outside the scope of this book.

#### 16.2.3 Uses of surface data

Surface data have a great many potential uses in many application areas. They are commonly used as relief maps to convey 3D information within the 2D environment and they are essential for 3D visualization. DEMs are also required for the ortho-rectification of optical images and for making terrain corrections to radar and gravity survey data. They also form a valuable data source for the calculation of descriptive surface parameters, for flow modelling (such as of water and mass movements), for geomorphological terrain analysis and for the calculation of other engineering, hydrological and hydrogeological indices.

Elevation models are used in applications over a growing range of scales, from global, meso and topo (conventional mapping scales) to micro and nano scales (very small-scale measurements), and within a diverse range of disciplines – meteorological, geological and geomorphological, engineering, biological and architectural. At the micro and nano scales, close-range photogrammetry is now a rapidly developing science. Examples in these fields include the micro-scale analysis of the terrain of stream beds, to reveal the physical characteristics of habitats occupied by small organisms such as fish and crustaceans, and the analysis of rock fracture surfaces for modelling fluid flow or of the frictional properties affecting rock strength.

### 16.3 Visualizing surfaces

GIS visualization tools, in both two and three dimensions, rely on the ability to share and integrate data, models and interpretations. The simplest form of visualization involves the display of 2D images with conventional cartographic symbols. For instance, the geological map can be recreated using conventional geological symbology, which can

easily be incorporated within GIS, through the use of special fonts. Thus structural information, for instance, can be presented to give the appearance of a published geological map.

GIS also bridges the gap between 2D and 3D displays (and analysis). This is especially useful in geosciences, because depth is such a fundamental consideration and the integration of sub-surface data has become an essential part of any digital mapping technology.

Visual exploration and interrogation, in several dimensions, facilitates enhanced understanding of structures and relationships. Virtual field visits become a routine activity and are vital in assisting in logistical planning and to improve understanding prior to setting foot in the field, thus saving valuable time and reducing risks. All GIS software provides 2D and 3D tools for the manipulation of surfaces, images and maps in pseudo 3D space and for the mathematical derivation of other products, such as slope angle, aspect or azimuth, curvature, line of sight, watersheds and catchments. Examples of software providing excellent 3D manipulation and viewing capabilities are ERDAS Imagine's Virtual GIS, ER Mapper, ArcGIS 3D Analyst, MapInfo Vertical Mapper and Geomatica Fly.

information but they are not visually powerful. A contour is a line connecting points of equal surface value. Contour lines reveal the rate of change in values across an area for spatially continuous phenomena. Where the lines are closer together, the change in values is more rapid. They are drawn at a specified interval, which represents the interval as simply the change in  $z$  value between the contour lines. For example, a contour map of precipitation with a contour interval of 10 mm would have contour lines at 10, 20, 30 and so on. Each point on a particular contour line has the same value, while a point between two contour lines has a value that is between the values of the lines on either side of it. The interval determines the number of lines that will be on a map and the distance between them. The smaller the interval, the more lines will be created on the map. A base contour may also be specified, as a starting point; this is not the minimum contour, but refers to a starting point from which contours go both above and below, based on the contour interval. For example, the base contour may be set to 0 and the interval may be set to 10. The resulting contour values would be -20, -10, 0, 10, 20 and 30. Watson (1992) provides a comprehensive description of the concepts of all forms of surface modelling, including algorithms for contouring and surface generation.

### 16.3.1 Visualizing in two dimensions

Appreciation of a 2D representation of the truly 3D physical landscape requires some level of interpretation and imagination but conventional flat maps do not make such appreciation very easy. Cartography has traditionally made use of a range of visual symbols to show height information and create the illusion of an undulating surface: elevation contours, spot height symbols, hill shading and cliff and slope symbols. GIS allows much more through the simulation of the 3D environment but we often still need to use 2D output to convey the results of our efforts. Again GIS cleverly provides us with the ability to visualize 3D quantities and objects within a 2D medium, through the use of *shaded relief* or *hillshades* and *contours*.

#### 16.3.1.1 Contours

Contour lines are the more familiar and mathematically more precise way of representing surface

#### 16.3.1.2 Shaded relief or hillshading

Hillshading is a technique used to create a realistic 2D view of terrain by simulating light falling on a surface from a given direction and simulating the shadows this creates. It is often used to produce visually appealing maps that are easier to interpret. Used as a background, hillshades provide a relief over which both raster data or vector data can be displayed.

There are three types of hillshading: *slope shading* where tonal intensity is proportional to the angle of slope (e.g. the steeper the slope, the darker the tone); *oblique light shading* where the pattern of light and dark on the surface is determined by a simulated oblique light source; and *combined shading* which represents the combination of these two types.

Contours and hillshading are quite often used together since they complement one another: hillshading provides a qualitative impression of the terrain, while contours show quantitative height information but only for discrete locations.

Oblique light shading involves the simulation of the oblique illumination of a surface by defining a position (angle and height) for an artificial parallel light source and calculating a brightness value for each position, based on its orientation (on the basis of slope and aspect) relative to the light source. The surface illumination is estimated as an 8 bit value in the range between 0 and 255, based on a given compass direction relative to the Sun and an altitude above the horizon. *Analytical oblique light hillshading* estimates the brightness based on the angle between the selected illumination direction (vector) and the surface normal vector (see Figure 16.2).

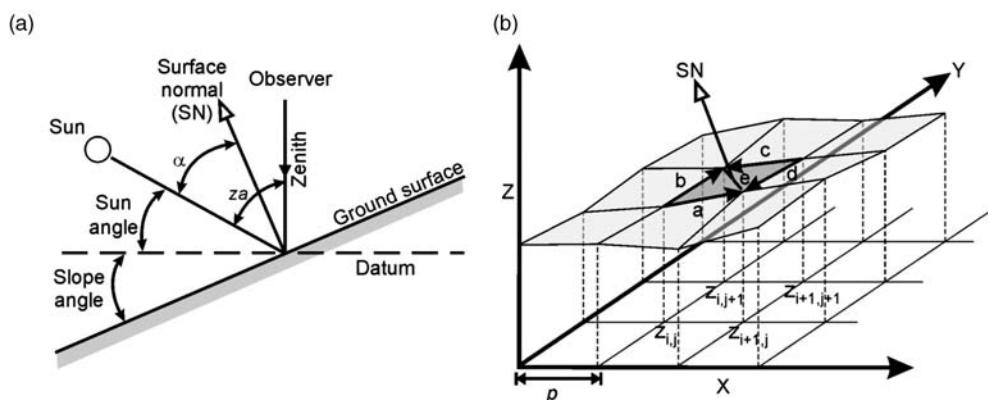
Conceptually, the illumination vector is defined by two angles, an angular attitude relative to north, or azimuth (given as a compass direction between  $0^\circ$  and  $360^\circ$ ), and an altitude (given as a horizontal angle between  $0^\circ$  and  $90^\circ$ ). The surface normal (SN) is a vector perpendicular to a surface, as defined in the raster case, by a pixel and its closest eight neighbouring grid cells (as illustrated in Figure 16.2b). The DN value assigned to the output pixel will be proportional to the cosine of the 3D angle between the surface normal and the illumination vector (shown as  $\alpha$  in Figure 16.2a) – a technique originally suggested by Wiechel (1878). So for slopes almost normal to the illumination direction, the angle will be very small, the cosine

value large and the estimated brightness proportionally high.

There are a number of ways of calculating this from a raster surface. The altitude and azimuth of the illumination source and the slope and aspect angles of the pixel being evaluated are needed. From these parameters, hillshade ( $h$ ) can be calculated as follows (where all angles are in radians):

$$h = 255[(\cos(z\alpha)(\cos(\text{slope}))) + (\sin(z\alpha)\sin(\text{slope})\cos(\text{az} - \text{aspect}))] \quad (16.1)$$

where  $z\alpha$  is the zenith angle and  $az$  the azimuth angle. Altitude is normally expressed in degrees above the horizontal but the formula requires the angle to be defined from the vertical, i.e. the zenith angle ( $z\alpha$ ) which is measured between the overhead zenith and the illumination direction (see Figure 16.2a). It represents the  $90^\circ$  complement of altitude. The azimuthal angle of the illumination must be changed from its compass bearing to a mathematical unit ( $maz$ ), i.e.  $maz = 360^\circ - az + 90^\circ$ , and if this angle is greater than or equal to  $360^\circ$ , then  $maz - 360^\circ$ . Note that formula (16.1) is essentially the same as (3.27) introduced in Section 3.7.1 for true Sun illumination.



**Figure 16.2** (a) Slope geometry for computation of hillshades and (b) illustration of the SN for a raster surface. SN is perpendicular to the average cross-product  $e$  of  $ab$  and  $cd$ , which approximates the surface area occupied by the pixel, depending on the resolution. Modified after (a) Reichenbach *et al.* (1993) and (b) Corripi (2003)

The  $x$ ,  $y$ ,  $z$  values of the vectors along the sides of the central pixel,  $a$ ,  $b$ ,  $c$  and  $d$ , in Figure 16.2b are given by

$$\begin{aligned} a &= (p, 0, \delta z_a), \text{ where } \delta z_a = z_{i+1,j} - z_{i,j} \\ b &= (0, p, \delta z_b), \text{ where } \delta z_b = z_{i,j+1} - z_{i,j+1} \\ c &= (-p, 0, \delta z_c), \text{ where } \delta z_c = z_{i,j+1} - z_{i+1,j+1} \\ d &= (0, -p, \delta z_d), \text{ where } \delta z_d = z_{i+1,j} - z_{i+1,j+1} \end{aligned} \quad (16.2)$$

The vector normal to the surface in the central pixel in Figure 16.2b is then defined by

$$\begin{aligned} \text{SN} &= \frac{a \times b}{2} + \frac{c \times d}{2} \\ &= \frac{1}{2} \begin{vmatrix} i & j & k \\ p & 0 & \delta z_a \\ 0 & p & \delta z_b \end{vmatrix} + \frac{1}{2} \begin{vmatrix} i & j & k \\ -p & 0 & \delta z_c \\ 0 & -p & \delta z_d \end{vmatrix}. \end{aligned} \quad (16.3)$$

Simplifying 16.3, we have the cell orientation defined by the heights of the central pixel corner points:

$$\text{SN} = \begin{pmatrix} 0.5p(z_{i,j} - z_{i+1,j} + z_{i,j+1} - z_{i+1,j+1}) \\ 0.5p(z_{i,j} + z_{i+1,j} - z_{i,j+1} - z_{i+1,j+1}) \\ p^2 \end{pmatrix}. \quad (16.4)$$

Once SN is derived, the 3D angle between the surface normal and the illumination vector can be

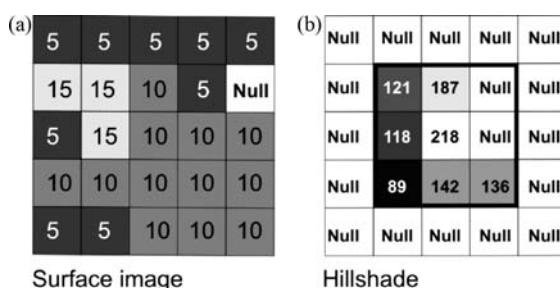
calculated and the DN value assigned to the output pixel will be proportional to  $\cos(\alpha)$ . A simple example is shown in Figure 16.3.

The values returned by hillshading may be considered a relative measure of the intensity of incident light on a slope. Such measures could be useful for many applications, such as selecting suitable sites for particular agricultural practices or slopes suitable for ski resorts. Hillshading of DEMs can also accentuate faults and other geological structures and facilitate geological interpretation. Convolution gradient filtering can also be used to identify linear features in remotely sensed imagery but the result may not be as visually striking as a shaded-relief image derived from a DEM.

When creating a cartographic hillshade for visualization, convention dictates that the light source is placed in the north-west (upper left) quadrant of the map, so casting a shadow at the bottom right of the object. The eye tends to see objects better when the shadow is cast at the bottom of the view; placing the light source elsewhere creates a visual effect that makes hills look like hollows.

### 16.3.2 Visualizing in three dimensions

Visualization in ‘pseudo 3D’ requires height to convey the third dimension, whether it is topography or some other attribute. In fact several things are needed: the definition of the 3D coordinate space, a viewing perspective, a vertical exaggeration to control depth and (optionally) a



**Figure 16.3** Simple raster example: (a) the input surface; and (b) the calculated hillshade, with  $135^\circ$  as the azimuth and  $45^\circ$  as the altitude. The bold line in (b) encloses the nine central pixels for which hillshade values can be calculated using the moving  $3 \times 3$  window; null values are generated around the edges and adjacent to any nulls in the input raster, hence the two null values inside the bold line

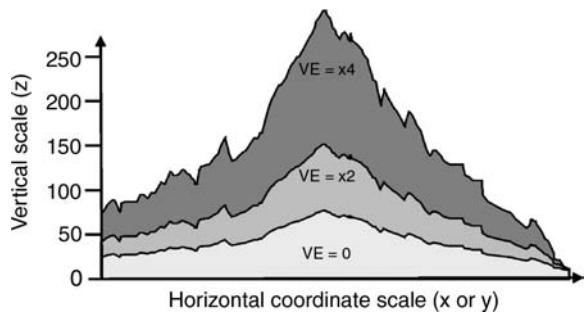
simulated light source, in addition to the datasets being visualized. Certain parameters necessary for 3D visualization are connected to the data (*base or ground elevations*), while others are temporary or virtual and control the environment of visualization (*artificial illumination* and *vertical exaggeration*).

#### 16.3.2.1 Basal or ground elevations

These refer to the basal or ground-level values used to display raster images or vector objects within the 3D space. They are needed to place an object correctly according to the  $z$  scale of the display. The height values can be derived from one of several sources: the values of the layer being displayed (the nodes of a TIN or the pixels' DN values of a DEM); the values stored in a different TIN or raster layer which covers the same geographic area; or lastly, from a value or expression (this would produce elevation at a fixed value or as a function of another attribute). Figure 16.4 illustrates the principle and effects of this.

#### 16.3.2.2 Vertical exaggeration

This refers to a kind of relative scaling used within 3D views to make subtle surface features (and any objects on the surface) more visible. This scaling applies to the environment of visualization rather than to the datasets being visualized; it is a temporary visual effect produced by multiplying the height values in the display by a constant factor. A vertical exaggeration of 2 multiplies all heights by 2 and exaggerates the vertical scale, whereas an



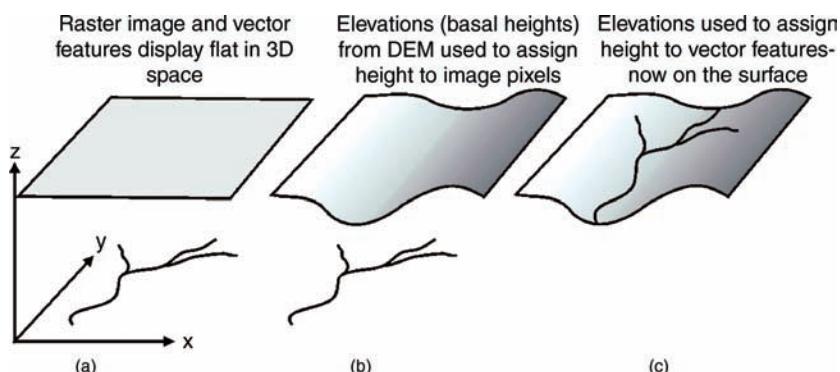
**Figure 16.5** Schematic illustration of vertical exaggeration (VE) in 3D environment. Three surface elevation profiles are shown, with no exaggeration (lowest profile), and with VE factors of  $\times 2$  and  $\times 4$  (middle and upper profiles respectively)

exaggeration of 0.1 multiplies all heights by 0.1 and suppresses the vertical scale, and so on, as shown in Figure 16.5.

Vertical exaggeration has two main uses: to emphasize subtle changes in elevation on a surface that is relatively flat or has great extent; and to force the  $x$ ,  $y$  units into proportion with the  $z$  units, if these represent different quantities or units.

#### 16.3.2.3 Projection of 2D vector objects into 3D space

Vector projection or ‘extrusion’ represents the 3D projection of 2D vector features. For example, an extruded point becomes a line; an extruded line becomes a wall; an extruded polygon becomes a block. Objects can also be projected downwards, below a surface, so that a point representing a well



**Figure 16.4** Schematic illustration of the effect of assigning basal elevations to a raster surface and vectors in 3D space: (a) flat raster and vector features; (b) surface extruded but vectors are still flat; and (c) both surface and vectors have height and the vectors now plot on the surface

or borehole could be extended to represent its depth below ground level. In contrast to the base elevations, extrusion controls the upper elevation of features and the simulation of 3D objects using 2D map features. Since they are simulations, feature extrusions can be said to produce ‘geotypical’ or generic representations of physical objects, rather than ‘geospecific’ ones (i.e. actual objects).

Features can be extruded by a variety of methods, as shown in Figure 16.6. A simple line feature is shown whose heights, when added to by various methods, are altered to simulate other features. By adding a constant value, the line feature is extruded upwards by a constant value to become a wall (50 units, as shown in Figure 16.6b); this can be applied to points and lines only. A second form of extrusion is formed when a value is added to the minimum or maximum height of the feature (base or top), and all other vertices are extruded to the same absolute value, whether up or down (as shown in Figure 16.6c and d); this can be applied to lines and polygons only. Lastly the vertices can be extruded to a specified absolute value, whether above or below the original values, as shown in Figure 16.6e; this can be applied to points, lines and polygons.

#### 16.3.2.4 Artificial illumination

This also applies to the environment of visualization rather than to the datasets themselves and is calculated by the same principle as the estimation of oblique light hillshading. Every 3D display has a theoretical light source, the position of which controls the lighting and shading of the display. The pseudo-illumination geometry is defined by azimuth and altitude angle settings. Again azimuth is a compass direction, measured clockwise in degrees from 0 (due north) to 360 (also due north).

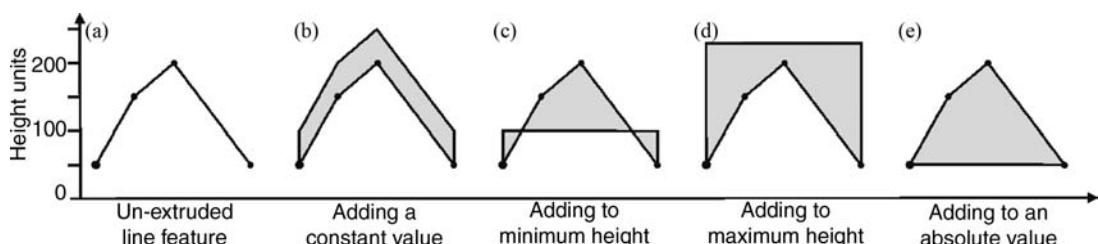
Altitude is the angle, measured in degrees from 0 to 90, between the light source and the horizon. An altitude of 0° means the light source is level with the horizon; an altitude of 90° means it is directly overhead.

## 16.4 Extracting surface parameters

Surface or terrain parameters, such as slope angle (gradient) and orientation (aspect), are important controls on a number of natural processes, such as rainfall runoff and erosion, and incident solar radiation upon slopes. There are many published texts describing different methods of calculating these parameters from DEMs. As described in detail in Chapter 4 and briefly in Chapter 14, the calculation of slope gradient, aspect and curvature, from a raster surface, are essentially neighbourhood operations or point spread functions, involving the use of a convolution kernel which is passed over the raster to produce a new set of values that describe the variance of each parameter and the morphology of that surface. Here we provide a summary of the parameters and the more common methods for their calculation, referring the reader to further texts where appropriate. Each parameter is illustrated here using a simple  $5 \times 5$  raster surface.

### 16.4.1 Slope: gradient and aspect

A slope is defined by a plane tangent to the surface, as modelled by a DEM at a point, and it confers the angle of inclination (steepness) on that part of the surface. While typically applied to topography, slope may be useful in analysing other phenomena,



**Figure 16.6** Illustration of the mechanisms of extruding 2D vector objects into the 3D perspective environment, for the purposes of visualization

for example for a surface of rainfall intensity, showing where intensity is changing and how quickly (steeper ‘slopes’ indicate values that are changing faster).

Slope has two component parameters: a quantity (*gradient*) and a direction (*aspect*). Gradient ( $g$ ) is defined as the maximum rate of change in altitude and aspect ( $a$ ) represents the compass direction, or azimuth, of this maximum rate of change. For the geoscientist this is equivalent to dip and dip direction of an inclined bedding or structural surface. More analytically, slope gradient at a point is the first derivative of elevation ( $z$ ) with respect to the surface slope, where  $g$  is the maximum angle, and the direction or bearing of that angle is the aspect. Since gradient has direction, it is a vector product. At the same time the first derivative at a point can be defined as the slope (angular coefficient or trigonometric tangent) of the tangent to the function at that particular point.

The general mathematical concept and calculation of gradient for raster data have already been introduced in Chapter 4 (see Section 4.4, formulae (4.12), (4.14), (4.15)) in relation to high-pass filters. Here we address the same parameter with direct relevance to surfaces and in particular to DEM data.

#### 16.4.1.1 Gradient

Slope gradient may be expressed as either degrees or per cent; the former are commonly used in scientific applications, while the latter is more commonly adopted in transport, engineering and other practical applications. Per cent gradient is calculated by dividing the elevation difference (known as the rise) between two points by the distance between them (known as the run), and then multiplying the result by 100. The degree of

gradient is derived from the geometric relationship between the rise and run, as sides of a right-angled triangle, the angle opposite the rise. Since degree of slope is equal to the tangent of the fraction of rise over run, it can also be calculated as the arctangent of rise over run. Measures of slope in degrees can approach 90° but measures in per cent can approach infinity, for instance in the case of a vertical cliff. An example is shown in Figure 16.7.

So for a raster grid, gradient is calculated, on a pixel-by-pixel basis, within a moving  $3 \times 3$  window, as the maximum rate of change in values between each pixel and its neighbours. The gradient ( $g$ ) is calculated using a second-order finite difference algorithm based on the four nearest neighbours:

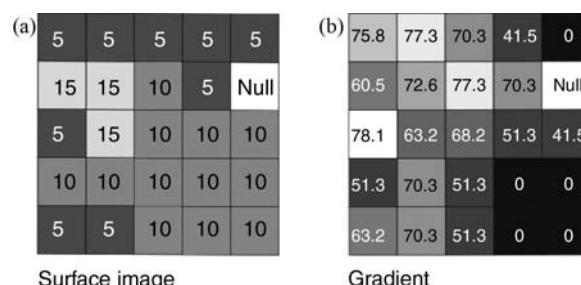
$$\tan g = \sqrt{[(\delta z / \delta x)^2 + (\delta z / \delta y)^2]}. \quad (16.5)$$

Such measures are given in degrees or per cent, according to taste:

$$g^{\text{degree}} = \arctan\left(\frac{g^{\text{per cent}}}{100}\right) \quad (16.6)$$

$$g^{\text{per cent}} = \tan(g^{\text{degree}}) \times 100. \quad (16.7)$$

Common slope procedures involve calculation from the pixel values immediately above, below, to the left and to the right of the central pixel, but not the corner (diagonal) pixel values, and in such cases this is referred to as the *rook’s case* because it resembles the way the rook moves on a chessboard. While elevation models are commonly stored as integer data (normally 16 bit), the output from a slope calculation will always be a real number, that is floating point.



**Figure 16.7** Simple raster example: (a) surface (integers); and (b) gradient (floating point)

Since  $g$  is usually calculated in radians, conversion to degrees is given by

$$g^{\text{degree}} = \arctan(g^{\text{radian}}) \times \left(\frac{180}{\pi}\right). \quad (16.8)$$

See Sections 20.4, 21.1 and 21.3 for examples of the use of slope (gradient) within the GIS case studies section in Part Three.

#### 16.4.1.2 Aspect (or azimuth)

The companion component of gradient, aspect ( $a$ ) identifies the *downslope orientation* or *direction of gradient*, measured with respect to north. When calculated from surface elevation (topography), it is usually referred to as ‘aspect’; in reference to other parameters the term ‘azimuth’ tends to be used.

The pixel DN values in a raster representing aspect are compass directions or bearings, in degrees, measured in a clockwise direction from 0 to 360, where north has a value of  $0^\circ$ , east  $90^\circ$ , south  $180^\circ$  and west  $270^\circ$ . An example is shown in Figure 16.8.

Aspect is calculated in the same ‘rook’s case’ manner, as in the calculation of gradient, as follows:

$$\tan a = \frac{(\delta z / \delta x)}{(\delta z / \delta y)}. \quad (16.9)$$

Any pixels calculated as having a zero slope, representing areas which are ‘flat’, are given a special aspect value, usually  $-1$ , to indicate that they have no aspect direction:

$$\text{if } (\delta z_x = 0 \text{ AND } \delta z_y = 0), a = -1. \quad (16.10)$$

An example of the use of aspect within a slope stability hazard assessment can be found in Section 21.3 in Part Three.

#### 16.4.2 Curvature

*Curvature* ( $c$ ) represents the rate of change in surface orientation of a variable across an area. It is calculated from a surface (raster or vector), such as elevation, and describes the convexity or concavity of that surface. Referring to Figure 16.2, it can be considered as a measure of the variation in the SN across the image or map, and is therefore the first derivative of the surface normal vector, and the second derivative of position on a surface (with respect to the changing rate of gradient  $g$ ; see also Section 4.4).

Several measures of curvature are recognized: *profile (downslope) curvature* and *cross-sectional (plan) curvature*, which are orthogonal to one another; and *total curvature* as a summation of profile curvature and cross-sectional curvature. Profile curvature is parallel to the direction of maximum gradient (or aspect) while cross-sectional curvature is perpendicular to it. Essentially, total curvature is the Laplacian, as further explained later.

Geomorphological forms can be discriminated in digital images by their curvature forms, examples of which are illustrated schematically in Figure 16.9; for example, ridges are convex in cross-section and valleys are concave in cross-section (where gradient and Laplacian may be variable in both cases), whereas planar slopes have zero cross-sectional curvatures since gradient is constant and thus curvature is zero. Peaks are convex in both cross-profile and cross-sectional (i.e. in all directions) and the reverse is true for pits, which are concave in all directions. If areas are flat or slopes are planar, curvature is zero.

As the summation of the changing rate of gradients in  $x$  and  $y$  directions for a 2D dataset,

(a)	5	5	5	5	5
	15	15	10	5	Null
	5	15	10	10	10
	10	10	10	10	10
	5	5	10	10	10

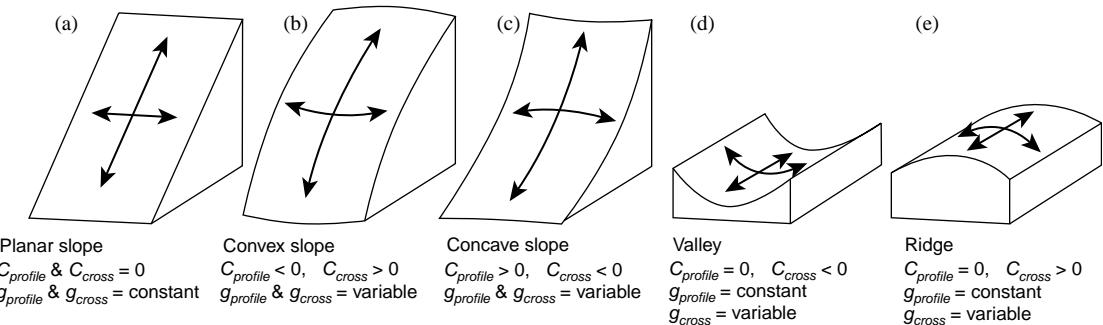
Surface image

(b)	198.4	171.9	153.4	135.0	-1.0
	135.0	168.7	135.0	153.4	Null
	293.2	341.6	90.0	180.0	225.0
	-1.0	333.4	-1.0	-1.0	-1.0
	341.6	333.4	270.0	-1.0	-1.0

Aspect

**Figure 16.8** Simple raster example (a) surface (integer); and (b) aspect grid (floating point)

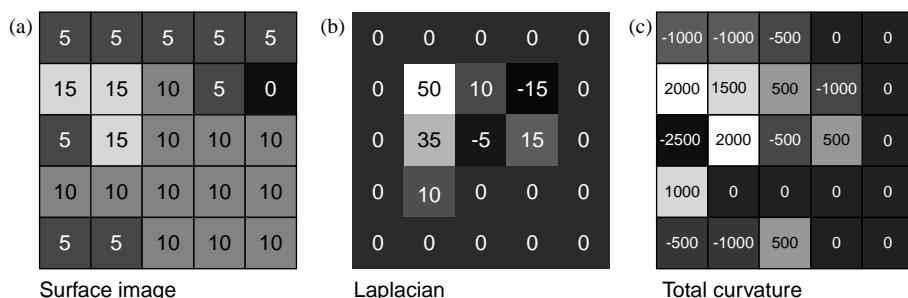


**Figure 16.9** Some common schematic slope forms and the relationships between curvature and gradient: (a) planar slope; (b) convex slope; (c) concave slope; (d) channel; and (e) ridge, where curvature may be positive (concave) or negative (convex)

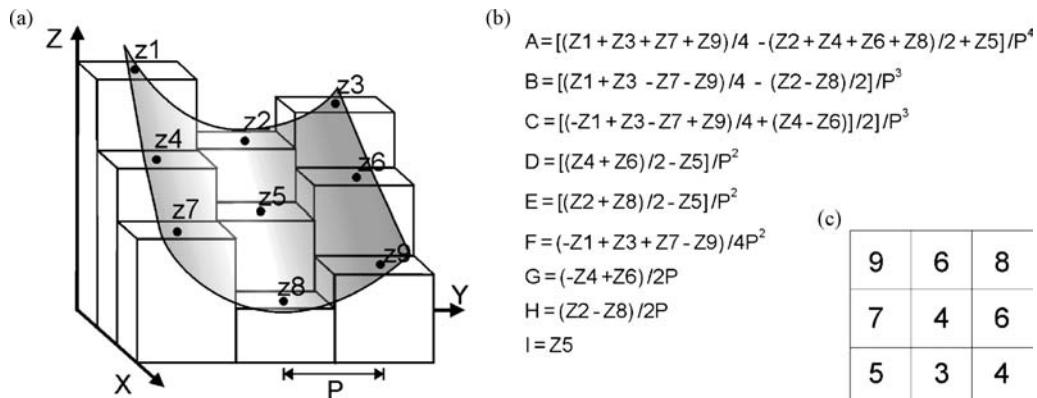
the Laplacian represents the curvature at a point without direction, i.e. total curvature, and it can be derived by the application of a Laplacian convolution filter to a raster surface (see Section 4.4). From differentiation theory, a negative Laplacian indicates a convex surface and a positive Laplacian a concave one. If, however, we recall the principle of the Laplacian filter in Section 4.4, we know that the conventional Laplacian filter in image processing gives the reverse so that a positive Laplacian indicates convexity and a negative one concavity.

The Laplacian, as a scalar, is composed of components in both  $x$  and  $y$  directions. We can decompose it in the aspect, or gradient, direction and in the direction perpendicular to aspect, by a second partial differentiation of elevation ( $z$ ) in these two directions. We denote these as  $c_{profile}$  and  $c_{cross}$ , and they represent the profile curvature

and cross-sectional curvature as introduced earlier. For simplicity, we denote the Laplacian as  $c_{total}$ . There is a convention, in some GIS software packages, that the three measures of curvature are calculated in such a way that positive values for  $c_{total}$ , or  $c_{cross}$ , indicate an upwardly convex surface, whereas a positive  $c_{profile}$  indicates an upwardly concave surface; the illustrations in Figure 16.9 follow this convention. A zero value for any of these indicates no curvature. Other packages give all forms of curvature with positive values to indicate convexity and negative ones for concavity. Regardless of any convention, the Laplacian, using a convolution kernel, is the simplest and reliable method of deriving curvature from raster data. A simple example is shown in Figure 16.10, with the Laplacian result in Figure 16.10b.



**Figure 16.10** Simple raster example: (a) surface; (b) the Laplacian, as representing total curvature (as would be calculated using the  $3 \times 3$  kernel, the outer rim of pixels always having zero value because of the kernel size); and (c) total curvature as calculated using the polynomial fitting method (as in Figure 16.11). Since both (b) and (c) represent total curvature, positive values indicate convexity and negative values indicate concavity. Both input and output are integer values



**Figure 16.11** (a) Pixel diagram illustrating the relation of raster elevations to a conceptual curved, channel-like (concave) surface. Modified after Zevenbergen and Thorne (1987). (b) Equations used to derive the various directional components of curvature from the surface in (a) where  $P$  is the increment between two pixels in either the  $x$  or  $y$  direction. (c) Sample elevation  $z$  values representing the feature shown in (a)

Mathematically, true curvature along the aspect direction  $\alpha$  is a function of both first and second derivatives defined as

$$c_{\text{true}} = \frac{(\partial^2 z / \partial \alpha^2)}{[1 + (\partial z / \partial \alpha)^2]^{3/2}}. \quad (16.11)$$

The estimation of curvature, as carried out in many proprietary software suites, including ESRI's ArcGIS and RiverTools, follows the method first formulated by Zevenbergen and Thorne (1987). This method estimates curvature using a second-order polynomial surface (a parabolic surface) fitted to the values in a  $3 \times 3$  window centred at  $x$  and  $y$ ; the surface is constructed in the manner illustrated by the block diagram in Figure 16.11a, using parameters in Figure 16.11b, and is of the general form as shown here:

$$c = Ax^2y^2 + Bx^2y + Cxy^2 + Dx^2 + Ey^2 + Fxy + Gx + Hy + I \quad (16.12)$$

where  $c$  is the curvature function at position  $x$  and  $y$  as illustrated in Figure 16.11a. Coefficients  $A$  to  $I$  are derived as in Figure 16.11b (the result is shown in Figure 16.10c). The methods used in Idrisi and Landserf also use polynomials but of slightly different form.

Referring to the values in the block diagram in Figure 16.10, the term  $c_{\text{profile}}$  can be estimated by (parameter  $D$  in Figure 16.11b)

$$c_{\text{profile}} = \frac{200(DG^2 + EH^2 + FGH)}{(G^2 + H^2)} \quad (16.13)$$

and  $c_{\text{cross}}$  as (parameter  $E$  in Figure 16.11b)

$$c_{\text{cross}} = \frac{-200(DH^2 + EG^2 - FGH)}{(G^2 + H^2)}. \quad (16.14)$$

The total curvature,  $c_{\text{total}}$ , is derived from  $D$  and  $E$  as in Figure 16.11b as

$$c_{\text{total}} = -200(D + E) \quad (16.15)$$

which for the values shown in Figure 16.11c, and  $p = 10$ , gives  $c_{\text{profile}} = 1.4$ ,  $c_{\text{cross}} = -4.6$  and  $c_{\text{total}} = -6$ , indicating that the feature represented by the values in this  $3 \times 3$  window is a channel, whose bed is slightly concave along its length, and that the overall window curvature is concave.

Several software packages employ this polynomial fitting method but there are some concerns with the derivation of curvature in this way. Firstly, the polynomial surface may not necessarily pass exactly through all nine elevation points ( $z_1 \dots z_9$ ). Secondly, if the complexity of the actual surface (represented by the  $3 \times 3$  grid in Figure 16.10a) is a plane, then the coefficients  $A$  to  $F$  will be zero. It must be stressed that this method does not represent the calculation of true curvature, merely directional estimates of it.

From an applied viewpoint, curvature could be used to describe the geomorphological characteristics of a drainage basin in an effort to understand erosion and runoff processes. The gradient affects the overall rate of movement downslope, aspect defines the flow direction and profile curvature

**Table 16.1** Geomorphological features, their surface characteristics and the pixel curvature formula relationships, where positive values of  $c_{\text{profile}}$  indicate upward concavity, and the opposite is true for  $c_{\text{total}}$  and  $c_{\text{cross}}$

Geomorphological feature	Surface characteristics	Second derivatives: profile and plan curvature
Peak	Point that lies on a local convexity in all directions (all neighbours lower)	$c_{\text{profile}} < 0, c_{\text{cross}} > 0$
Pit	Point that lies in a local concavity in all directions (all neighbours higher)	$c_{\text{profile}} > 0, c_{\text{cross}} < 0$
Ridge	Point that lies on a local convexity that is orthogonal to a line with no convexity/concavity	$c_{\text{profile}} \approx 0, c_{\text{cross}} > 0$
Channel	Point that lies in a local concavity that is orthogonal to a line with no concavity/convexity	$c_{\text{profile}} \approx 0, c_{\text{cross}} < 0$
Pass	Point that lies on a local convexity that is orthogonal to a local concavity (saddle)	$c_{\text{profile}} < 0, c_{\text{cross}} < 0$
Plane	Points that do not lie on any surface concavity or convexity (flat or planar inclined)	$c_{\text{profile}} = 0, c_{\text{cross}} = 0$

determines the acceleration (or deceleration) of flow. Profile curvature has an effect on surface erosion and deposition and as such is a very useful parameter in a variety of applications. Cross-sectional curvature affects convergence and divergence of flow, into and out of drainage basins, and so can be used to estimate potential recharge and lag times in basin throughput (see also Section 16.4.3).

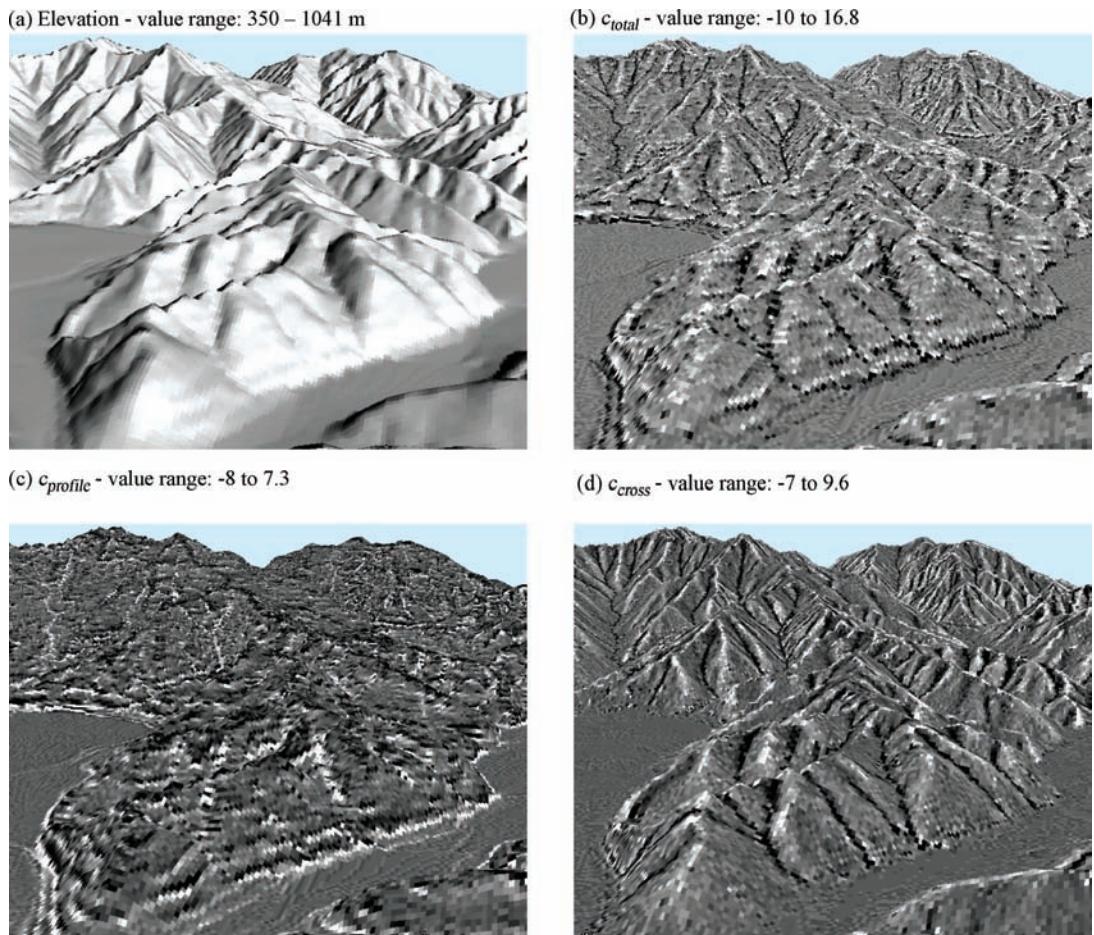
The shape of the surface can be evaluated to identify various categories of geomorphology, such as ridges, peaks, channels and pits. These are summarized in Table 16.1, and illustrated in the 3D perspective terrain views shown in Figure 16.12.

#### 16.4.3 Surface topology: drainage networks and watersheds

Several other very important parameters can be extracted from surfaces, and in particular DEMs, and these relate to the connection between geomorphological features and the fluvial processes which produce them, hence the term

‘surface topology’. An increasing number of tools are now available for the extraction of stream networks, drainage basins, watersheds and flow grids from raw elevation data. These combine calculations of aspect and curvature to establish the movement of water across a surface, by defining connectivity between drainage channels and to construct catchments and flow networks. This represents an area extensive research and one to which we cannot do justice here.

The extraction of drainage networks has been investigated by many authors and these methods now appear in many proprietary software suites (e.g. RiverTools, ArcGIS’s hydrological toolbox and Idrisi). There are many tricky issues associated with automated drainage network extraction from DEMs and it has been suggested that these are caused by scale dependence in the methods and a failure to accommodate the fractal natures of both elevation and drainage. Some algorithms begin with the calculation of flow direction or flow routing, then creation of drainage basins from their outlet points, which are ‘chased’ progressively upwards to the drainage divides (e.g. River Tools). Potential problems occur where

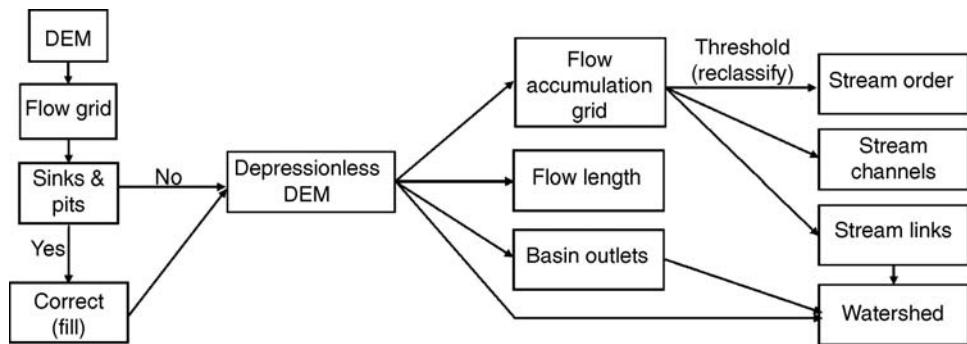


**Figure 16.12** Raster surfaces representing: (a) elevation; (b) total curvature ( $c_{total}$ ); (d) profile curvature ( $c_{profile}$ ); and (d) cross-sectional or plan curvature ( $c_{cross}$ ). A grey-scale colour lookup table with low values in black and high values in white is used in each case. As calculated using ArcGIS's Spatial Analyst curvature calculator, where positive values of  $c_{profile}$  indicate upward concavity, whereas the opposite is true for  $c_{total}$  and  $c_{cross}$

there are pits in the DEM surface since these tend to stop the ‘drainage chasing’ algorithms, and so these should be corrected first. Any pits in the DEM must be filled to create a ‘depressionless’ DEM and this is used to derive flow direction and flow accumulation (to give a measure of flow through each pixel); the latter is then thresholded to give stream order which, combined with the calculated watershed, enables the drainage network to be derived (e.g. as in ArcGIS and illustrated in Figures 16.13 and 16.14). Geomorphological features, including drainage networks, can also be extracted by skeletonization of the DEM (see also Chapter 14).

#### 16.4.4 Viewshed

A *viewshed* represents an area or areas that are visible from a static vantage point. In the context of surface data, the viewshed identifies the pixels of an input raster (or positions on a TIN) which can be seen from one or more vantage points (or lines). The output product is a raster in which every pixel is assigned a value indicating the number of vantage points from which a pixel is visible. Visibility, in this respect, refers to the line of sight, but this ‘sight’ could also refer to the transmission of other signals, such as radio and microwaves. The viewshed is sometimes referred to as a 2D



**Figure 16.13** Schematic illustration of a drainage network extraction method (as used in ArcGIS)

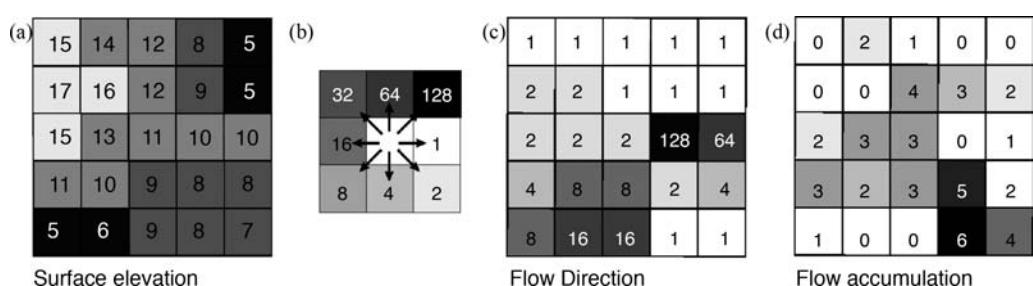
*isovist* (see online resources). Viewsheds are used in a variety of applications such as for the siting of infrastructure, waste disposal or landfill sites, and to select sites for transmission towers to avoid gaps in reception.

In an example with one vantage point, each pixel that can be seen from the vantage point is assigned a value of 1, while all pixels that cannot be seen from the vantage point are assigned a value of 0. The output is therefore a typically binary image.

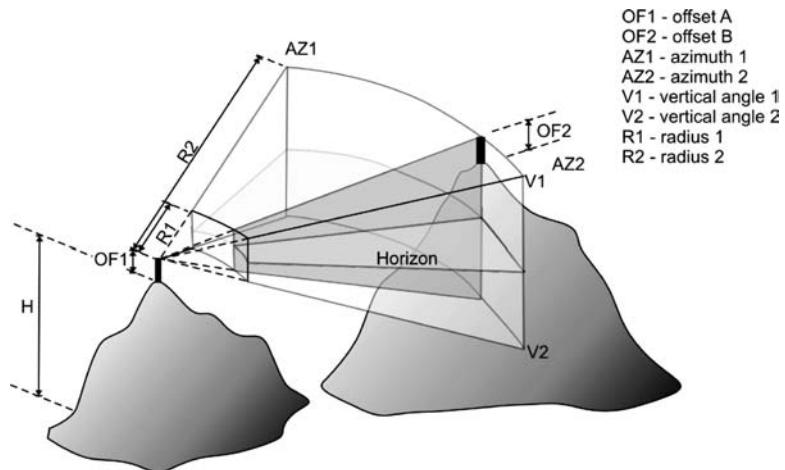
The simplest viewshed calculations assume that light travels in straight lines in a Euclidean manner, i.e. the Earth is not curved and no refraction of light occurs, and that there are no restrictions on the distance and directions of view. This assumption is acceptable over short distances (of several kilometres) but corrections for the Earth's curvature and optical refraction by the Earth's atmosphere are necessary for accurate results over longer distances.

To describe the concept and framework of a viewshed, several controlling parameters can be defined. These are the surface elevations of the vantage points, the limits of the horizontal angle to within which the viewshed will be calculated (given as two azimuths), the upper and lower angles, and the inner and outer radii (minimum and maximum distances from the vantage point), limiting the search distance within which the viewshed will be calculated from each vantage point, the vertical distance (if necessary) to be added to the vantage points, and the vertical distance (if necessary) to be added to the elevation of each pixel as it is considered for visibility. These are illustrated in Figure 16.15 and an example is shown in Figure 16.16.

A modification of this could be used to model areas where noise can and cannot be detected, such as from military installations or road traffic, known as a *soundshed*. Such soundshed analysis could



**Figure 16.14** Derivation of flow direction (or flow routing) and accumulation: (a) ‘depressionless’ input surface; (b) direction encoding kernel; (c) flow direction grid (each pixel is coded according to the direction in which flow would pass from it, according to the values in the kernel, always moving to the lowest of the adjacent pixels); and (d) flow accumulation grid (the number of pixels that flow into each pixel)



**Figure 16.15** Schematic illustration of parameters defining the viewshed: elevations of the vantage point(s) (H), the limits of the horizontal angle to within which the viewshed will be calculated (azimuths 1 and 2), the upper and lower view angles (V1 and V2) and the inner and outer radii (R1 and R2), limiting the search distance within which the viewshed will be calculated from each vantage point, the vertical distance (OF1 and OF2) to be added to the vantage points. Modified after the ESRI online knowledge base 2008

then prove useful in designing sound barriers around potentially ‘noise-polluting’ activities.

#### 16.4.5 Calculating volume

There may be many instances where we would like to estimate the volume of a quantity, as well as area or some other statistic. This may be very useful in the estimation of necessary costs associated with particular activities, such as for engineering or construction purposes, for example how much soil and rock need to be excavated to construct a cutting for a railway line, or the expected volume of a

reservoir. Such calculations are usually referred to as *cut and fill analysis*. Here we extract the change between two surfaces (usually raster). These input surfaces might represent the same location but recorded at different times. The calculation is made simple using raster because of the constant area of the pixel and the height given by the pixel DN, so that multiplying the two gives the volume occupied by that pixel. So that for a  $100 \times 100$  raster, of 10 m spatial resolution (i.e. an area of  $100 \text{ m}^2$ ), where each pixel DN represents a height of 2 m (above datum), the volume of the raster would be  $100 \text{ m} \times 2 \text{ m} \times 10000 \text{ pixels}$ , i.e. 2 million cubic metres. If for each pixel we calculate the volume, as

(a)	<table border="1" style="border-collapse: collapse; width: 100px; height: 100px;"> <tr><td>5</td><td>5</td><td>5</td><td>5</td><td>5</td></tr> <tr><td>15</td><td>15</td><td>10</td><td>5</td><td>Null</td></tr> <tr><td>5</td><td>15</td><td>10</td><td>10</td><td>10</td></tr> <tr><td>10</td><td>10</td><td>10</td><td>10</td><td>10</td></tr> <tr><td>5</td><td>5</td><td>10</td><td>10</td><td>10</td></tr> </table>	5	5	5	5	5	15	15	10	5	Null	5	15	10	10	10	10	10	10	10	10	5	5	10	10	10	Surface image	<table border="1" style="border-collapse: collapse; width: 100px; height: 100px;"> <tr><td>1</td><td>1</td><td>1</td><td>1</td><td>1</td></tr> <tr><td>1</td><td>1</td><td>1</td><td>1</td><td>Null</td></tr> <tr><td>0</td><td>1</td><td>1</td><td>1</td><td>1</td></tr> <tr><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td></tr> <tr><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td></tr> </table>	1	1	1	1	1	1	1	1	1	Null	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	Viewshed
5	5	5	5	5																																																		
15	15	10	5	Null																																																		
5	15	10	10	10																																																		
10	10	10	10	10																																																		
5	5	10	10	10																																																		
1	1	1	1	1																																																		
1	1	1	1	Null																																																		
0	1	1	1	1																																																		
0	0	0	0	0																																																		
0	0	0	0	0																																																		

**Figure 16.16** Simple raster example: (a) surface; and (b) viewshed (with simple vantage point indicated by a black cross, to which there are no added offset or specified angles or radii). Both input and output are integer values. Viewshed values of 1 or 0 indicate that the position is either visible or invisible (respectively) from the vantage point

a local operation, then one volume raster can be subtracted from or added to another to derive change. Depending on whether the change between two rasters is negative or positive, we have cut and fill statistics respectively.

## 16.5 Summary

It is worth noting once more the effect of the fractal nature of surface features, a commonly overlooked feature and one which has been commented on by many authors. Gradient, aspect and curvature are all phenomena which vary at different scales of topography, and so scales of observation. Measuring any of these phenomena in the field yields very different results according to the distance over which they are being measured. It has also been shown that the effect of spatial resolution on their calculation from raster imagery can have a profound effect on the result.

Nevertheless, surfaces are clearly a valuable data source from which a great variety of morphological descriptive information can be extracted, and for many different applications. Image processing techniques and GIS-based tools overlap significantly

in this chapter, though they are applied here with different intentions to those applied to multi-spectral images.

## Questions

- 16.1 What are the differences between continuously sampled and discrete statistical surfaces?
- 16.2 For any particular phenomenon you are working with, which type of surface description should you choose and why?
- 16.3 What are DEMs and why are they so powerful for geoscientific use?
- 16.4 What effect does spatial resolution have on the estimation of slope gradient from a raster surface?
- 16.5 How can parameters like slope and aspect be used to derive neighbourhoods? How could such neighbourhoods then be used?
- 16.6 What other applications could there be for surface parameters such as gradient, aspect and curvature? That is, in application to surface data other than topography.



# 17

## Decision Support and Uncertainty

### 17.1 Introduction

Uncertainty in GIS is inevitable and it arises for many reasons. The data we have at our disposal are never complete, and our knowledge and understanding of a problem are flawed or limited, because of natural variation, because of measurement error or because the information is out of date. Albert Einstein is famously quoted as having stated that ‘as far as the laws of mathematics refer to reality, they are not certain; as far as they are certain, they do not refer to reality’. We cannot get away from it or ignore its existence; we therefore must learn to live with uncertainty and deal with it.

While we realize that we may not be able to tackle directly the causes or sources of risk and uncertainty, we can recognize the existence of the uncertainty and attempt to quantify it, track its course through any analysis and estimate or predict its probable effect on the outcome. The more advanced and involved our spatial analysis becomes, the more input factors are aggregated and the greater effect any potential errors will have. One other important thing that GIS allows us to do in this respect is to simulate or model potential outcomes and scenarios, varying the input parameters and the effects of errors and uncertainties as we go. These processes act as a form of quality control or validation for both data and analysis. Through this process, not only do we better understand the nature of the errors and uncertainties, but

we also improve our understanding of the problem and the potential reliability of the result, by more closely defining the limits of its applicability. Some of the key publications covering this subject include those by Goodchild and Gopal (1989), Burrough and Frank (1996), Burrough and McDonnell (1998) and Foody and Atkinson (2002).

Three key concepts, which perhaps require definition or clarification in this context, are *decision support*, *uncertainty* and *risk* (and hazard). We will attempt to explain what we mean by these terms and how they are relevant to GIS. This chapter attempts to describe some of the surrounding issues and causes of and potential solutions to the problem of uncertainty.

### 17.2 Decision support

A *spatial decision support system* (SDSS) can be thought of as a knowledge-based information system which supports decision making or, more simply, is a mechanism to bring parameters together. An SDSS could involve a system designed to assist managers and/or engineers where the task at hand may be complex and where the aim is to facilitate skilled judgement. An SDSS could also be used to assist in problems which are of a poorly understood nature, or where data are incomplete, or where there are variables of unknown significance involved.

Indeed there are many definitions because the SDSS is used in many, very different contexts.

A ‘decision’ should be based on the level of acceptable risk and on the degree of confidence (error and uncertainty) in the available data. A decision may also imply the need for a quantitative prediction which demands the evaluation of the influential criteria, to which the decision rules will be applied. A good decision may lead to a bad outcome (and vice versa) but if good decisions are persistently attempted, then good outcomes should become more probable. Such decision making can be subdivided according to the situations in which the decisions are made, as follows:

- *Deterministic decision making:* This occurs when the ‘controls’ on the problem and the data are understood with some degree of certainty; so too are the relationships between each decision and the outcome. In such cases, categorical classes, rules and thresholds can be applied.
- *Probabilistic decision making:* Here the surrounding environment, relationships and outcomes are uncertain (to some degree). In general, this approach treats uncertainty as ‘randomness’ but this is not always the case, and especially not in the natural environment. Since a probabilistic approach tends to produce only a true or false result, degrees of uncertainty can only be accommodated if it is considered as a separate and distinct state.
- *Fuzzy decision making:* This approach deals with uncertainties which are related to natural variation, imprecision, lack of understanding or insufficient data (or all these). Such ambiguities can be accounted for by allowing that classes can exist in varying amounts rather than as one of two end-member states (true or false) so that an infinite number of further states, representing the increasing *possibility* of being true, can be accommodated.

Probability and possibility form opposing but complementary concepts which coexist within Dempster–Shafer theory, described in Chapter 18.

Examples of applications in which the SDSS is frequently used might include, for example, a classification of locations in an area according to their

estimated suitability for a pipeline route, or for a landfill site, toxic waste disposal or a hazard assessment. Within these and other applications, the function of the SDSS is to help decision maker(s) to identify areas where there are unacceptable levels of risk associated with various predictive outcomes, so that they can then select appropriate courses of action.

## 17.3 Uncertainty

In general terms, uncertainty can be considered as an indeterminacy of spatial, attribute or temporal information of some kind. It can be reduced by acquiring more information and/or improving the quality of that information. There will be few cases where it can be removed altogether, so it needs to be reduced to a level tolerable to the decision maker. Methods for reducing uncertainty include defining and standardizing technical procedures, improving education and training (to improve awareness), collecting data more rigorously, increasing spatial/temporal data resolution during data collection, field checking of observations, better data processing methods and models, and developing an understanding of error propagation in the algorithms used.

Assumptions must be made in all spatial analyses where any kind of ‘unknowns’ or uncertainties need to be dealt with. Examples of assumptions include that soil and rock classes (or any classes) are homogeneous across the area they represent; that slope angles classified as stable are stable everywhere; that classifications made at the time of data collection have not changed since then; or that geological boundaries are rigid and their positions are certain, everywhere. Just as uncertainties are unavoidable, so too are these assumptions. There are methods that we can employ to quantify these uncertainties and so limit the effect of the assumptions, such as allowing for the gradational nature of natural and artificial boundaries and for establishing threshold values for classification and standardization. Uncertainties are many and complex, and the underlying rule, once again, is to know the limitations of and to understand the data from the start.

Conditions of ‘certainty’, in contrast, could include situations where there is only one ‘state of

nature' or where any 'state of nature' that exists has only one effect on the outcome or only one outcome. Clearly such definitive, certain and simplistic states are rare or unlikely in nature but they are useful concepts from which to consider more realistic possibilities. There may be cases where an element of certainty may be acceptable, perhaps with respect to either data availability or cost, since both are often in short supply. Such shortages often lead to compromise, when some areas of uncertainty may have to be ignored.

Uncertainties within spatial analysis may be related to the validity of the information itself (*criterion* uncertainties), to the potential effects of the phenomena (*threshold* uncertainties), or to the handling of the information (*decision rule* uncertainties).

### 17.3.1 Criterion uncertainty

*Criterion uncertainty* arises from errors in original measurement, identification and/or data quality (perhaps during data collection by a third party). Broadly speaking, criterion uncertainty may be considered to be related to measurement (primary data collection) or conceptual (interpretative). It includes locational, attribute value, attribute class separation and attribute boundary uncertainties and they may not be correctable. In such cases an important step is to estimate or record the potential errors for the benefit of future users, and for your own liability. Measurement errors may derive directly from instruments with limited precision or as a result of user error, observer bias, mismatches in data collected by different individuals, sampling errors or poor sampling. Understanding the target and objective is vital in designing the sampling strategy. Repeated sampling can often correct for such potential errors but the expense of doing this may be prohibitive.

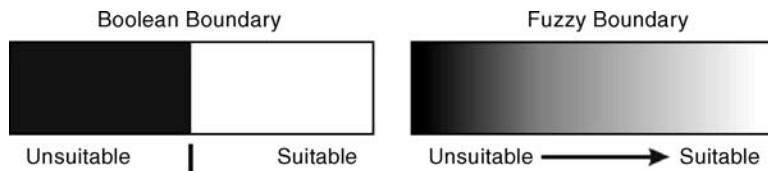
Criteria prepared for spatial analysis inherently include some uncertainty (error) since they have probably been rescaled or converted to discrete rasters from continuously sampled rasters, reclassified and/or generalized; they are thus the product of subjective decision making. Errors in spatial data are usually considered to be normally distributed and identifying them requires some ground truth information to allow comparison of the differences

between measured and observed, such as through the RMS error calculation. One thing is common to all: the more criteria that are involved in the decision making, the greater the uncertainty will be. In this situation, one could restrict the analysis to a simplistic combination of a few criteria but the simplistic solution will still incorporate uncertainties; it will merely ignore them since it cannot account for them. In the end there are two choices where data or criterion uncertainty is concerned: to reject the data (in favour of better data) or accept them and work around the uncertainties they contain. In many cases, the latter is the only course of action since there are no 'better' data.

### 17.3.2 Threshold uncertainty

There are two principal causes of *threshold uncertainty*. Firstly, the phenomena we describe are generally heterogeneous, that is we choose homogeneous classes for simplicity and convenience. The concept of 'possibility' could therefore be very useful when attempting to define class boundaries in natural phenomena, i.e. when deciding whether an object belongs to one class or another. Secondly, the boundaries between natural phenomena of any kind are rarely rigid or Boolean in character, because again we define arbitrary classes for our own convenience (they may not exist in reality). If we are able to treat such divisions less rigidly, we can in effect blur the boundaries between them. This will allow further possible states to exist; that is, in addition to 'suitable' or 'unsuitable', or prospective and non-prospective, stable and unstable values can be incorporated which represent the increasing likelihood of belonging to a class or state, as illustrated in Figure 17.1.

Similarly, considerable research has been carried out into the use of multi-source datasets to generate mineral 'prospectivity' estimates. In many cases, identifying the areas of very low and very high prospectivity has not been so difficult but uncertainty arises in describing the areas of intermediate prospectivity, which then require further analysis and interpretation. In such cases prospectivity (suitability) should be treated as a continuous phenomenon in representing a measure of confidence in an outcome.



**Figure 17.1** Different boundary types between class thresholds: (a) crisp (Boolean) threshold producing categorical classes of unsuitable and suitable; and (b) a fuzzy threshold representing increasing probability of membership of the class ‘suitable’

### 17.3.3 Decision rule uncertainty

*Decision rule uncertainty* refers to the way in which we apply thresholds to particular criteria to denote values of significance. If we have firm and reliable evidence about some phenomenon we may be able to apply ‘hard’ deterministic decision rules confidently. There may be many cases where some prescriptive law governs the analytical selections we make, in which case we also have to apply hard decision rules in order that the result complies with that law. Conversely, where we have incomplete data and must rely on circumstantial evidence then we should find a way to apply ‘soft’, probabilistic or fuzzy decision rules, ones in which a certain degree of confidence/uncertainty is attached to each result.

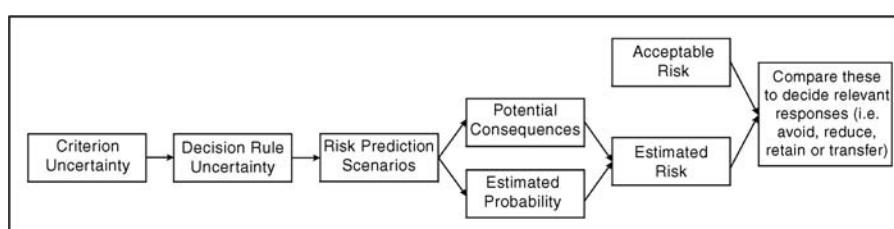
This type of uncertainty arises because of subjective judgements made by the decision maker in the ranking and weighting of criteria, in the method of criteria combination and in the choice of alternative courses of action. Spatial interpolation also falls into this category, since the nature and accuracy of the result are directly affected by the interpolation method chosen, and there is no unambiguously correct method. Uncertainties of this type are the most difficult to quantify, since they are not simply caused by mistakes or imprecision, and they may never be precisely known.

## 17.4 Risk and hazard

The existence of uncertainties and errors in both data and spatial analysis introduces an element of risk. Risk can be thought of not only as the prediction of the likelihood of a (potentially unwelcome) event occurring, but also as the chance of getting that prediction wrong. It is therefore normal for such spatial analysis to form part of a risk assessment framework of some kind, an example of which is illustrated in Figure 17.2.

There are many definitions of risk and there is a tendency for terms like risk and hazard, risk and uncertainty, or risk and probability to be used interchangeably, but in all cases this may be because of a misunderstanding of their meanings. In many senses, *hazard* represents the probability of the occurrence of an unwelcome event and *risk* is that probability modulated by the economic value of the losses per event. To this end, Varney (1984) defined the following relationship in an attempt to separate the related terms of risk and hazard and *vulnerability*, which represents a measure of the economic value (i.e. damage and cost, in both economic and human senses):

$$\text{Risk} = \text{Hazard} \times \text{Vulnerability}.$$



**Figure 17.2** Generalized risk assessment framework

Risk then represents the expected degree of loss, in terms of probability and cost, as caused by an event. Within the context of uncertainty, risk can be described by three phases of activity: (i) the prediction of one or more adverse events (scenarios) which have unexpected or unknown outcomes; (ii) the probability of those predicted events occurring; and (iii) their actual consequences. In defining risk, we should also define an *acceptable level of risk*, since it is a variable quantity. One popular method is that proposed by Melchers (initially in 1993, later published in 2000), which is known as the ‘as low as reasonably practicable’ or ALARP principle. This represents the minimum limit below which risk can be practically ignored. This is subjective and there are several alternative definitions, such as the ‘lower limit beyond which further risk reduction cannot be justified’. A common response to risk associated with uncertainty is to reduce the risk, either by *risk retention* (bearing the consequences) or *risk transfer* (i.e. insurance). However, all potential outcomes might not be insurable!

accepted values. The standard deviation is usually taken as a measure of accuracy and it is normally stated with reference to an interval, that is  $\pm$  a given value. High precision necessarily does not indicate high accuracy, nor does high accuracy imply high precision.

Errors are either systematic or random. *Random errors* occur when repeated measurements do not agree and they tend to be normally distributed about the mean. *Systematic errors* tend to be either positively or negatively skewed, indicating that they are of common source and of similar sign and value, such as an instrument error. Errors may be of position or attribute, and their source could be measurement (imprecision or instrument malfunction) or interpretation (conceptual). Ground-based sampling is usually taken to provide data which represent ‘true’ values. The errors are then identified by comparison with the ‘true’ values at the sample locations. For these locations, the *root mean square errors* (RMSE) can be calculated as

$$\text{RMS} = \left[ \frac{\sum_i (x_i - x_{it})^2}{n-1} \right]^{0.5} \quad (17.1)$$

where  $x_i$  is the measured value,  $x_{it}$  is the true value and  $n$  is the number of measurements (and hence error values). The RMS is commonly used for error description in GIS for continuously sampled data. In classified or categorical data, errors are usually described using a confusion matrix (see Section 8.5.2) which is constructed by cross-tabulation of observed (true) and mapped (estimated) values. This type of error assessment is commonly applied to classifications made from remotely sensed images, and has been the driving force for a great deal of work on validation within remote sensing applications. In such cases, the errors may come from any of a great many sources: misregistration of images, sensor properties, classification errors, ground truth data errors, class definitions, pixel impurities and more. Such errors are commonly irregularly distributed (spatially) but the confusion matrix technique does not reveal this pattern. Alternatively a geostatistical approach could be used to model the geographical variation of accuracy within the results. Accuracy assessment in classifications from remotely sensed data are

## 17.5 Dealing with uncertainty in spatial analysis

There are a number of tools we can employ to quantify, track and test for ambiguity within our data and analytical methods. These include error assessment, fuzzy membership functions, multi-criteria evaluation, error propagation and sensitivity analysis.

### 17.5.1 Error assessment (criterion uncertainty)

*Errors* can be defined as the deviation of a data value from the considered or measured ‘true’ value. Other terms that need defining here include *precision* and *accuracy* since these vary within some interval according to errors in the data. Precision can be described as the level of measurement ‘exactness’, and is usually limited by the instrument and/or method, whereas accuracy may be thought of as the degree to which information matches true or

therefore not trivial but are now considered to be fundamental to all forms of thematic mapping. Despite this, there still seems to be no universally accepted standard method of doing so; refer back to Section 8.5.2.

### 17.5.2 Fuzzy membership (threshold uncertainty)

Fuzzy logic is commonly applied in one of two basic ways, either through the combination of Boolean-type maps, using fuzzy rules, to yield a fuzzy output map, or through the use of fuzzy membership functions to rescale the data for submission to further analysis of varying kinds. The latter incorporates a measure not only of a phenomenon that exists on the ground, but also of some level of knowledge (confidence) about that phenomenon.

*Fuzzy membership* or *fuzzy sets* provide an elegant solution to the problem of threshold and decision rule uncertainty by allowing ‘soft’ thresholds and decisions to be made. Fuzzy membership removes the requirement of ‘total membership’: instead of just two states of belonging for a class, a fuzzy variable can have one of an infinite number of states ranging from 0 (non-membership) to 1 (complete membership) and the values in between represent the increasing possibility of membership. A simple linear fuzzy set is defined by

$$\mu(x) = \begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & a < x < b \\ 1 & x > b \end{cases} \quad (17.2)$$

where  $a$  and  $b$  are upper and lower threshold values of  $x$  defining the significant limits for the fuzzy set. Each value of  $x$  is associated with a value of  $\mu_{(x)}$  and ordered pairs  $[x, \mu_{(x)}]$ , and these together comprise the fuzzy set.

The fuzzy set membership function can be most readily appreciated with reference to a simple linear function but it may actually be linear, sigmoidal or J-shaped, and monotonic or symmetric in form (Figure 17.3). The threshold values which define it will depend on the phenomenon and desired outcome of the operation; the threshold values applied to each membership function reflect their

significance on the result but the function is not necessarily linear. Here we have only presented the linear fuzzy set membership function to illustrate the principle. The resultant fuzzy set layers can be combined in a number of ways, for example using Boolean logic and fuzzy algebra or ‘set theory’.

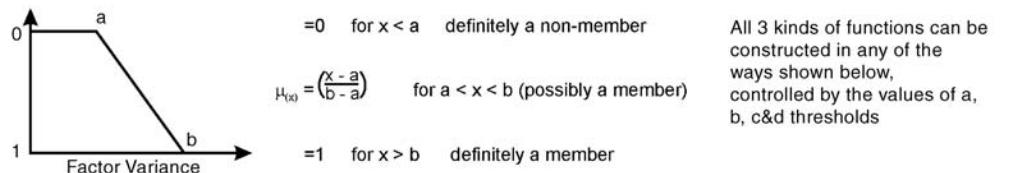
An illustration of multi-criteria evaluation applied to hazard assessment, using fuzzy scaled inputs (see also Section 15.5.6), is described in a research case study in Section 21.3.

### 17.5.3 Multi-criteria decision making (decision rule uncertainty)

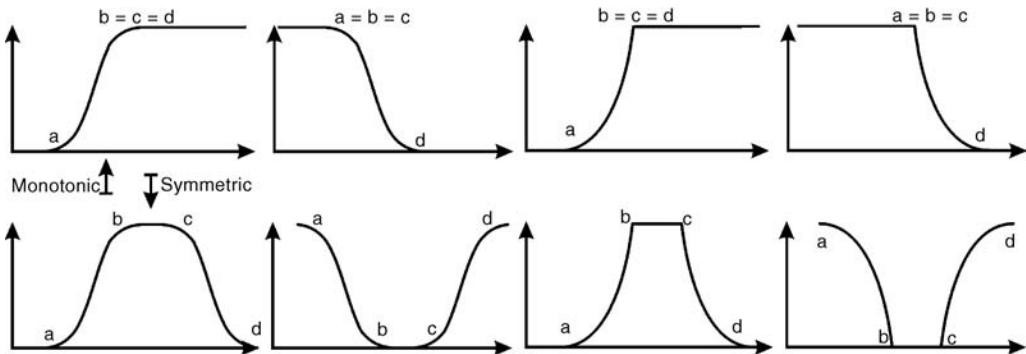
In this context we introduce the concept of *multiple criteria evaluation* or *multi-criteria decision making*, a process by which the most suitable areas for a particular objective are identified, using a variety of evidence layers, and any conflicts between objectives are resolved. The risks arising from decision rule uncertainty here are reduced through the integration of multiple criteria representing the various contributory processes and quantities. This process allows the input criteria to be handled in different ways according to their desired function within the analysis. This is further enhanced by the evaluation of individual criteria significance, the ranking of those criteria and the assignment of weights to give certain criteria variable influence inside the model. The criteria being evaluated are generally one of two types: *constraints* which are inherently Boolean and limit the area within which the phenomenon is feasible, or *factors* which are variable and are measured on a relative scale representing a variable degree of likelihood for the occurrence of the phenomenon. Decision rules, applied to these criteria, are commonly based on a linear combination of factors together, along with the constraints, producing an *index of suitability* or *favourability*. (These topics are discussed further in the next chapter.)

It is worth noting here the difference between *multi-attribute decision making* and *multi-objective decision making*, since both fall into this category of activities, and, further to this, the possibility of involving either individual and multiple decision making. Multi-criteria decision making is often used to cover both types, multi-attribute and multi-

### 1. Linear Membership Functions



### 2. Sigmoidal Membership Functions



**Figure 17.3** Fuzzy set membership functions: linear, sigmoidal, J-shaped, monotonic and symmetric

objective. The *attribute* is a measured quantity or quality of a phenomenon, whereas an *objective* infers the desired state of a system being analysed or simulated. The difference between individual and multiple decision making is not really about the number of people making the decisions but about the number of objectives being satisfied. For instance, an individual decision-making process with one single goal may involve a group of several people or just one, and a multiple decision-making process, involving a coalition of participating individuals, may be competitive or independent, depending on the nature of the topic. Group decision making is common in the public sector, especially since stakeholder involvement is often mandatory and there may be many points of view to be considered.

Commonly used multiple criteria combination methods which incorporate measures of uncertainty include weighted factors in linear combination (WLC), weights of evidence modelling (Bayesian probability), vectorial fuzzy modelling and Dempster–Shafer analysis (belief and plausibility). These will be described in more detail in the next chapter.

### 17.5.4 Error propagation and sensitivity analysis (decision rule uncertainty)

#### 17.5.4.1 Error Propagation

This refers to the process of determining the expected variability in the result as produced either by known errors, or by errors deliberately introduced into the input dataset. There are several popular methods of error propagation, including Monte Carlo simulation (which is probably the most widely used) and the analytical method or Taylor's series error propagation analysis.

#### 17.5.4.2 Monte carlo simulation

The Monte Carlo method is one of a group of rather computer-intensive general methods for assessing the impact of statistical errors on the results of functions. For any particular variable  $z$  which is a function of various inputs  $i_1 \dots i_n$ , the idea is to determine the error associated with the estimated value of  $z$ , and the contributions of each input to the error. The variable is considered to have a normal probability distribution function with known mean and variance, and stationarity is assumed. The function is calculated iteratively, to generate values

of  $z$  many times, so as to derive the average resulting  $z$  value and its standard deviation.

This method is commonly used to assess the errors associated with the calculation of surface parameters from a DEM, such as slope and aspect. In this context, for an error dataset with a mean of  $\mu = 0$ , variance  $\sigma^2$  is produced for a standard deviation of  $\pm 1$  m for instance; this is added to the DEM. Slope, for instance, is then calculated many times, for example 100 times, to produce 100 slightly different results. These are combined to produce an average slope image. Division of the standard deviation by the average slope then gives a measure of the relative error, in both magnitude and distribution.

#### 17.5.4.3 Analytical method

This method uses a mathematical (polynomial-type) function to describe the way errors are translated through a particular decision rule. If a function has continuous derivatives (up to  $n + 1$ th order) it can be expanded and if continued expansion causes the function to converge, then it is known as a Taylor's series, which is infinitely differentiable. Error propagation of this type involves evaluating the effect of known errors on the function, where only the lowest order terms are considered important since any pattern can be simulated if a function of high enough order is used. These methods are described in detail by Heuvelink, Burrough and Stein (1989), Goodchild and Gopal (1989) and Burrough and McDonnel (1998).

#### 17.5.4.4 Sensitivity analysis

This process revolves around the view that the attribute values and their weights, within spatial analysis, are the most important aspect since they comprise the most subjective part of the analysis. The scaling and weighting of datasets involves interpretation and perceptive judgements, which introduce 'error'; if the rank order of inputs changes greatly as the weights are modified, then the latter should be re-evaluated. If the order does not change, then the model can be considered quite stable. Detailed descriptions of sensitivity analysis methods are provided in several texts listed in the references.

The main difference between this and error propagation methods is that sensitivity analysis requires

prior understanding of data and errors, whereas error propagation is a process by which errors are introduced into the analysis. Sensitivity analysis involves a number of methods aimed at determining the degree to which the result is affected by changes in the input data and decision rules (and weights); it is a measure of the robustness of a model.

#### 17.5.5 Result validation (decision rule uncertainty)

Multi-criteria analysis is often carried out with little consideration of the meaning, relevance or quality of the output solution or the effect of potential errors it contains. The multi-criteria evaluation procedures discussed here and in the next chapter provide a means to allow for, quantify and reduce certain varieties of uncertainty, but it is in the final stages more than any other that the validity of the result should be questioned and tested. This often means some kind of 'blind' test to determine the validity of the output suitability map. To do this, some reliable 'ground truth' information is required. This could come from a physical ground test but this is in itself subjective. What is really needed is an objective measure of the effectiveness of the result, revealing how predictive it is.

There are several methods for *cross-validating* or estimating the success of a particular result using some *training data*. This refers to the use of a partitioned dataset, with one part used in the analysis and the other retained for confirming or testing the result. The ground truth data are usually provided by some known occurrence data, such as known landslide locations or mineral occurrences. A widely used approach utilizes the confusion matrix which we encountered in Section 8.5.2.

Several other error measures are used, which are based on a *pairwise comparison* approach, such as the *kappa statistic*, which describes the agreement between measured and observed spatial patterns, on a scale between 0 and 1, and is described formally in Section 8.5.2 and simply as follows:

$$\kappa = \frac{p(a) - p(e)}{1 - p(e)} \quad (17.3)$$

where  $p(a)$  represents the relative observed agreement among the input values and  $p(e)$  represents the

probability that any agreement is caused by chance. If there is complete agreement then  $\kappa = 1$ ; if the opposite is true and there is no agreement then  $\kappa \leq 0$ . Cohen's kappa statistic is only applicable in cases where two inputs are compared. A variation of this can be used to consider multiple inputs. In Section 8.5.2, we have already presented the kappa coefficient derived from the confusion matrix for classification accuracy assessment with an example of multiple classes.

Other methods include the holdout or test-set,  $k$ -fold and leave-one-out cross-validation (LOOCV) methods. The *holdout or test-set method* involves a random selection of, for example, 30% of the data which are kept back from the analysis to act as test data. If data are not plentiful this may be unacceptable; we may not wish to waste 30% of the input data. The  *$k$ -fold cross-validation method* is slightly less wasteful and involves the random division of the dataset into  $k$  subsets, where a regression is performed  $k$  times on the subsets. Each time one subset is used as the test set and the other  $k - 1$  subsets are used as the training data, then the average error from the  $k$  regressions is derived. The *leave-one-out method* involves the iterative removal of one test data point, which is equivalent to the  $k$ -fold method taken to the extreme, where  $k = n$  ( $n$  being the number of test data points). This last method is useful if you have an independent set of very few ground truth data points. The entire analysis can be run iteratively, each time with one test data point removed, and the success of prediction examined each time. In this way, prediction curves can be constructed to give an idea of the effectiveness of the model and data used.

## 17.6 Summary

Uncertainty is an area of very active research within spatial analysis and remote sensing since the volume of data and our access to it are both growing. Clearly, what we do as geospatial scientists with our digital data in GIS is fraught with dangers and vague possibilities. These problems are often, alarmingly, overlooked. They cannot be avoided but there are things we can do to minimise the risks and allow for the uncertainties and errors. Whether we are becoming more or less critical of data quality and reliability is a moot point but it is certain that we should continue to develop tools and understanding to keep pace with these trends.

## Questions

- 17.1 List some examples of phenomena that cannot be realistically described by rigid (Boolean) functions and describe some fuzzy alternatives for each.
- 17.2 Why is error tracking important?
- 17.3 Why is it important to quantify uncertainty in GIS?
- 17.4 What are the main types of uncertainty and how do they affect the analysis?
- 17.5 What developments should GIS software provide in future to help deal with uncertainty?
- 17.6 Why is validation important in multi-criteria evaluation problems?
- 17.7 What can standards and benchmarks contribute?
- 17.8 How significant are metadata in this context and why?



# 18

## Complex Problems and Multi-Criteria Evaluation

### 18.1 Introduction

This branch of GIS activity is sometimes referred to as ‘advanced spatial analysis’, a term that tends to make it sound more complicated than it actually is. Certainly, it tends not to be the kind of everyday activity carried out by the mainstream of GIS users but is popular within the realm of the geosciences since it provides an elegant mechanism for tackling the complex processes of nature. Multi-criteria evaluation is in itself a topic of growing development and is sometimes considered not strictly part of GIS but an area of overlap with it, in which there is potential for shared gain. We observe that the use of multi-criteria decision analysis has grown considerably over the past 10–15 years, resulting in what is now a well-established body of research on this topic. Unsurprisingly, the use of the simplest methods greatly outnumbers that of the more complex procedures.

Any procedure that uses spatial data to satisfy a particular request could be described as ‘spatial analysis’ and often is. This chapter deals with the procedures by which we deal with complex geospatial problems to which there may be many, potentially unknown, contributing processes and pieces of evidence. These procedures incorporate the conversion of real-world problems into a set of

abstract quantities and operations, the accommodation of the vagaries of the natural environment, and of ‘unknowns’ quantities, to produce a realistic and practical solution to the original problem. The terms ‘criteria’ and ‘factor’ tend to be used interchangeably in this chapter in reference to the multiple input layers; however, ‘factor’ is used only in reference to continuously sampled (variable) data, whereas ‘criteria’ is used in a more general sense to both categorical and variable inputs.

Data generated from modern-day surveys and exploration campaigns are not only diverse but voluminous. Sophisticated topographic, geological, geochemical, remote sensing, geophysical (high-resolution ground and airborne) surveys not only make the analysis more quantitative (hopefully) but also make interpretation more difficult. A successful result lies in the effective processing of the data, extraction of the relevant factors and integration of these factors into a single ‘suitability’ map or index.

Over the past decade or so, many techniques have evolved to exploit large datasets and construct maps that illustrate, for example, how mineral potential or prospectivity changes over an area (Knox-Robinson and Wyborn, 1997; Chung, Fabbri and Chi, 2002; Chung and Keating, 2002), or how slope instability (or vulnerability to slope failure) varies across an area (Chung and Fabbri, 2003, 2005; Wadge, 1988;

Mason and Rosenbaum, 2002; Liu *et al.*, 2004a, to name but a few), or how rapid surface soil erosion can be discriminated from other kinds of small-scale surface change (Liu *et al.*, 2004b). These kinds of analyses (often referred to collectively as ‘modelling’ even though it really is not) demand the abstraction of reality, i.e. the representation of physical properties numerically, and the application of statistical approaches to accommodate natural variations.

In this chapter, we describe the main approaches and point to other more detailed texts, where appropriate. Generally speaking, and whatever the application area, the steps involved follow a similar path, always beginning with the definition of the problem, through data preparation to the production of a result and its validation, followed by some recommendations for action ‘on the ground’. These phases are illustrated in Figure 18.1.

## 18.2 Different approaches and models

There are a number of different approaches to multi-criteria decision making and analysis, with the aim of estimating ‘suitability’ or ‘favourability’ across a region, some of which pre-date GIS. They are often divided into two broad categories: the *knowledge-driven approach* (conceptual) and *data-driven approach*, and of the latter there are two further kinds. The first is empirical and tries to identify significant spatial relationships, and the second uses artificial

intelligence and neural networks objectively to recognize patterns in the data.

### 18.2.1 Knowledge-driven approach (conceptual)

This approach generally involves a specific model for an individual case or area. The ‘model’ is then broken down to its constituent parts to identify the significant contributing criteria. It is a method commonly used for hazard mapping but also in mineral exploration, especially in its early stages, perhaps on a regional scale or to identify areas for more detailed work. A database is then constructed which contains data appropriate to the description of the criteria. The criteria are then combined in such a way as to identify areas of potential. *Dempster–Shafer theory (DST)* (Section 18.5.5) and the *analytical hierarchy process (AHP)* (Section 18.5.6) fall into this category.

### 18.2.2 Data-driven approach (empirical)

This approach is most commonly applied to mineral prospectivity mapping where it exploits pre-existing knowledge of a particular type of mineral deposit, how it relates to its surroundings and its mode of formation. In this type of intuitive approach, the aim is to predict areas which are geologically similar to other known mineral deposits but which have not yet undergone any systematic exploration.

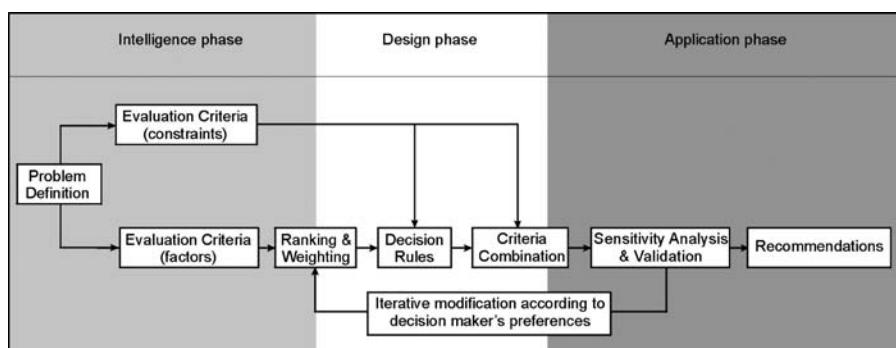


Figure 18.1 Conceptual framework for multi-criteria spatial analysis

For this, and the knowledge-driven approach, there are usually three steps involved: the identification of relationships, the quantification of those relationships as layers and then the integration of the layers. Research in this area has tended to concentrate on the third step and, as a result, there are an increasing number of techniques available, ranging from the very simple (Boolean set theory) to rather more complex algebraic, weights of evidence and fuzzy logic methods.

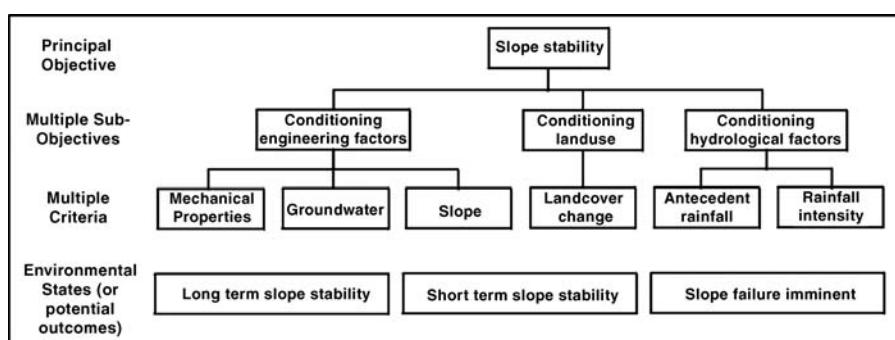
### 18.2.3 Data-driven approach (neural network)

The application of neural networks and other data-mining techniques involve the process of ‘learning’ and ‘pattern’ recognition from the data. Unlike the knowledge-driven and empirical data-driven approaches, the neural network approach evaluates all inputs simultaneously by comparison with a *training dataset*. This approach is particularly suited to very large data volumes where the number of significant factors and possible combinations is potentially huge, so the individual handling of layers and spatial relationships becomes impractical. Success relies heavily on the training or learning process and so, once trained for a particular case, the system must be retrained before it can be applied to any other case. In general, this approach is very demanding computationally and generally less commonly applied than the previous two, so we will not attempt to cover it in this book.

## 18.3 Evaluation criteria

Criteria are usually evaluated within a hierarchical structure, some consideration of which is necessary before the data are prepared. The overall objective of the work and any subordinate objectives should be identified (see Figure 18.2) before the requisite criteria may be identified to achieve the objectives. Once the hierarchy of objectives and criteria is established, each criterion should be represented by a map or layer in the database. Each chosen criterion should ideally be independent and unambiguous. Each must also be represented in common units, otherwise, when the layers are combined, the value units of the result will be meaningless.

Let’s consider as an example the problem of slope stability assessment, as a principal objective, beneath which are several sub-objectives (as illustrated in Figure 18.2). The latter may represent intermediary steps in achieving the principal objective or valuable end products in their own right. Each of the sub-objectives demands multiple input criteria to satisfy their purpose, for example three layers representing the variable mechanical properties of rocks and soils, groundwater levels and slope gradient may be required to produce a layer representing *conditioning engineering factors*, as a *factor of safety* measure. All the sub-objective layers may then be combined to produce the overall slope stability hazard assessment map. This map may then in turn be used to give predictions of the potential environmental states, for example if the area is stable, stable only in the short term or unstable.



**Figure 18.2** Schematic illustration of a multi-objective, multi-criteria evaluation system, under the main objective heading of slope stability analysis and with several possible outcomes (or environmental states)

An alternative example could be the evaluation of mineral prospectivity (as the principal objective) where the sub-objectives for the same geographic area could be evaluation of the potential for several materials of economic value, such as gold, nickel and base metals. Each sub-objective has a set of specific input criteria and while these may be used by more than one sub-objective, they may be prepared, reclassified and rescaled uniquely for each sub-objective. In each case, the end result is a map showing the variation in prospectivity (or suitability for exploration) for a given commodity over a given area. Thus we can see the distinction between multi-objective and multi-criteria decision making. The identification of both the significant input criteria and the various sub-objectives is therefore an important step. Following this the appropriate data must be identified and prepared, to represent these criteria.

In choosing the criteria, there are two problematic situations: too many criteria, so that the decision-making process becomes too complicated to understand; or too few criteria, causing an oversimplification of the problem. Poor understanding may cause the former and the latter is usually caused by data shortage.

There are a number of methods for selecting criteria, such as by researching past cases or by conducting a survey of opinions. As mentioned in the previous chapter, input criteria are generally of two types: continuous (later referred to as factors) and thematic (later referred to as constraints in some situations). The continuous type represent spatially variable, continuously sampled phenomena, often containing values on interval, ratio and cycle scales, such as those described in Chapter 12. The thematic type comprise representations of discretely sample phenomena, usually containing values on the nominal or ordinal measurement scale (as described in Chapter 12). The input criteria must be prepared and scaled, using operations and procedures described in Chapter 14, in such a way as to contribute correctly towards the end result.

## 18.4 Deriving weighting coefficients

In the process of establishing the structure of the model, it generally becomes clear that some criteria

play a more significant role than others in leading to the outcome. Having identified and prepared the input criteria, the next step is to assess and quantify their relative significance. To achieve this, the criteria must be ordered and a mechanism identified to describe the order numerically; using *ranking* and *weighting* procedures. Deciding on ranking and weighting is perhaps the most difficult aspect of multi-criteria evaluation problems, and it commonly requires discussion, field verification and iterative modification.

There are many weight derivation approaches and these differ in their complexity, accuracy and method. Weights should not be considered as simple indicators of criterion significance because they should allow for changes in the range of factor values as well in as the significance of each factor. The reason for this is that a factor weight could give an artificially small or large effect on an outcome simply by increasing or decreasing its range of values, for example weights in the range of 1–1000 will have a far greater effect than those in the range 1–10. Weights applied to the criteria should always sum to 1, so that

$$\sum w_i = 1 \quad (18.1)$$

where there are  $n$  criteria and where the weights ( $w$ ) range over  $w_1, w_2, w_3, \dots, w_n$ . Most multi-objective and multi-criteria evaluation procedures and decision-making processes involve the combination of a series of input variables, and it is highly likely that these inputs will contribute to the outcome to varying degrees. If their significance in contributing to the outcome is not equal, then some means of quantifying those contributions is necessary. Many types of weighting procedures have been proposed to allow this, and these include rating, ranking, pairwise comparison and trade-off analysis, some of which are more popular than others.

### 18.4.1 Rating

*Rating* involves the assignment of values on relative scales of significance, for example 0 to 10 or 100. One popular method, referred to as *point allocation*, involves the identification of a number of points or scores among the input criteria. For instance, if a 0–100 scale is chosen and there are

three input criteria, a value (*score*) of 60 out of 100 could be assigned to the most significant criterion, 30 to the next and 10 to the last. The resultant weights would then become 0.6, 0.3 and 0.1. Alternatively, in the *ratio estimation* method, scores are assigned to the most and least significant criteria; the latter is then used as the reference from which all other ratio scores are calculated. Criteria are ranked and scored on a relative scale, as in the point allocation method, and then the score of the least significant one is divided by each other criterion score to give a weight, which is then normalized by the sum of weights. The process is repeated for the second least significant, the third least significant, and so on until all criteria have calculated ratio weight values. The result is then a measure of the difference between lowest and highest values for a particular criterion in comparison with those of the first (least significant) criterion.

#### 18.4.2 Ranking

Here the criteria are first arranged in *rank order* according to their considered relative significance in affecting the outcome. The weights are then derived by one of a number of popular methods, summarized as follows. In all methods, the weights are normalized by the sum of the weights for all criteria:

1. *Rank sum*: This refers to the normalized summation of weights derived for each criterion, as follows:

$$w_i = \frac{n - r_j + 1}{\sum(n - r_k + 1)} \quad (18.2)$$

where  $w_i$  is the normalized weight for the  $i$ th criterion,  $n$  is the number of criteria being evaluated,  $r_j$  is the rank position of the  $i$ th criterion, and the criteria being evaluated ( $k$ ) range over  $k = 1, 2, 3, \dots, n$ .

2. *Rank reciprocal*: This involves calculation of the weight reciprocals, normalized by the sum of weights:

$$w_i = \frac{1/r_j}{\sum(1/r_k)}. \quad (18.3)$$

3. *Rank exponent*: In this case, a ‘most significant’ criterion is identified and a variable is set to represent it. It is then used as a power with which to multiply the normalization:

$$w_i = \frac{(n - r_j + 1)^p}{\sum(n - r_k + 1)^p}. \quad (18.4)$$

A weight is specified for the most significant criterion ( $p$ ). The value of  $p$  is then solved iteratively and the weights are derived. The higher the value of  $p$ , the more sharply the values of the normalized weights rise with increasing significance. If  $p = 0$  the weights will be equal in value; if  $p = 1$  the result is equivalent to the rank sum of weights. So this method allows a certain amount of control over or *trade-off* between the weights.

These three methods involve only relative weight ‘approximation’ so that the larger the number of criteria, the less appropriate the method becomes. They are therefore considered acceptable for cases with few input criteria (Table 18.1).

#### 18.4.3 Pairwise comparison

The pairwise comparison matrix (PCM) method was created and developed by Saaty (1980) for use within the analytical hierarchy process (described in Section 18.5.5). The method has received criticism for its abstraction from the real measured or reference scales of the input criteria. It is therefore vital that the input criteria are normalized correctly and to common scales before combination. The method is, however, flexible, easy to understand (since only two criteria are considered at a time) and appropriate for collective and iterative discussions of weighting. This method is incorporated into the decision support section of the Idrisi software suite. There are three steps involved, and these are summarized as follows and illustrated using slope stability assessment as an example:

1. *Construction of the PCM*: A matrix is constructed where every input criterion is compared with every other and is given a score representing its significance in contributing to the outcome.

**Table 18.1** Resultant weights derived using the rank sum, rank reciprocal and rank exponent methods above, for a set of ( $k$ ) criteria

$k$ ( $k$ )	Rank ( $r$ )	Rank sum		Rank reciprocal		Rank exponent	
		Weight ( $n - r_j + 1$ )	Normalized weight	Reciprocal weight ( $1/r_j$ )	Normalized weight	Weight ( $n - r_j + 1)^p$ , $p = 0.8$	Normalized weight
1	3	2	0.200	0.333	0.160	1.741	0.213
2	4	1	0.100	0.250	0.120	1.000	0.122
3	1	4	0.400	1.000	0.480	3.031	0.371
4	2	3	0.300	0.500	0.240	2.408	0.294
Totals		10	1.000	2.083	1.000	8.181	1.000

The values in the matrix are assigned from a relative scale of importance between 1 (equal importance) and 9 (extreme importance). Reciprocal values can be used to indicate the reverse relationship, for example 1/9 indicating that one factor is extremely less important than another. The values on the diagonal are always 1, where identical criteria are compared, and the values in the upper right part of the matrix are reciprocals of those in the lower left part. The assigned value scale is described in Table 18.2 and an example is given in Table 18.3.

2. *Derivation of weights:* The weights are produced from the principal eigenvectors of the PCM and are derived by hand using the following method. The values in each column of the matrix are summed, to give column marginal totals. A second matrix is then generated by dividing each matrix value by its column marginal total. These values are then averaged across the rows to derive the weight for each criterion.
3. *Calculation of a consistency index, within the matrix:* This ensures that the logical relationships between the criteria are represented fairly. The value of this ratio should be as low as possible, indicating that the relative comparisons have been made sensibly. This process involves the calculation of several component parameters as follows:
  - (a) *Weighted sum vector (WSV)* – where the first weight ( $w$ ) is multiplied by the first column value ( $ct$ ) in the matrix, the second by the second column, and so on. These values are then summed over the rows to give the WSV:
$$\text{WSV} = \sum_{i=1}^n wct. \quad (18.5)$$
  - (b) *Consistency vector (CV)* – here the WSV is divided by the criterion weights.
  - (c) *Average consistency vector ( $\lambda$ )* – this is calculated for all the criteria.
  - (d) *Consistency index (CI)* – since there are always inconsistencies within the matrix,  $\lambda$  is always greater than or equal to the number of input criteria ( $\lambda \geq n$ ) for any reciprocal matrix. The closer the value of  $\lambda$  to  $n$  ( $\lambda = n$  in an ideal case), the more consistent the matrix. So  $\lambda - n$  represents a

**Table 18.2** Table of significance estimations based on a nine-point scale. After Saaty (1980). Reciprocal values can also be used (see Table 18.3 below)

Significance	Value	Significance	Value
Extreme importance	9	Moderate to strong importance	4
Very to extreme importance	8	Moderate importance	3
Very strong importance	7	Equal to moderate importance	2
Strong to very strong importance	6	Equal importance	1
Strong importance	5		

**Table 18.3** Pairwise comparison matrix, used to assess relative factor significance, in contributing to slope instability, and to calculate criterion weights as shown in Table 18.4. Note that the table should be read from the left, along the rows, so that slope is considered the most significant and distance from drainage the least significant

PCM step 1	Slope	Aspect	Factor of safety	Distance from drainage
Slope	1	3	5	7
Aspect	1/3	1	1	7
Factor of safety	1/5	1	1	7
Distance from drainage	1/7	1/7	1/7	1
Marginal totals	1.68	5.14	7.14	22.00

good measure of consistency and CI as an estimate of the average difference gives a good judgement of consistency:

$$CI = \frac{\lambda - n}{n - 1}. \quad (18.6)$$

Let's consider the example of slope stability, where there are four input criteria considered to have varying degrees of influence in causing slope failure for an area. A PCM could be constructed and used to derive the criterion weights, using the method described, as illustrated in Tables 18.3 and 18.4.

Using these scores and weights, the WSV and CI can be derived for each criterion, as shown in Table 18.5.

For this example, with four criteria contributing to slope instability, the average consistency vector ( $\lambda$ ) is

$$\lambda = \frac{4.486 + 4.234 + 4.291 + 4.059}{4} = 4.268 \quad (18.7)$$

and so the calculated value of CI is

$$CI = \frac{4.268 - 4}{4 - 1} = 0.08921. \quad (18.8)$$

**Table 18.4** Second table generated from the column totals of those in Table 18.3 to derive the weights

PCM step 2	Slope	Aspect	Factor of safety	Distance from drainage	Weight
Slope	0.597	0.583	0.700	0.318	0.550
Aspect	0.199	0.194	0.140	0.318	0.213
Factor of safety	0.119	0.194	0.140	0.318	0.193
Distance from drainage	0.085	0.028	0.020	0.045	0.045
Marginal totals	1.000	1.000	1.000	1.000	1.000

**Table 18.5** Derivation of the consistency index (CI) using the values in Table 18.3 and the weights derived in Table 18.4

PCM step 3	Weighted sum vector (WSV)	Consistency vector (CV)
Slope	$(0.55)(1) + (0.213)(3) + (0.193)(5) + (0.045)(7) = 2.465$	4.486
Aspect	$(0.55)(0.33) + (0.213)(1) + (0.193)(1) + (0.045)(7) = 0.901$	4.234
Factor of safety	$(0.55)(0.2) + (0.213)(1) + (0.193)(1) + (0.045)(7) = 0.828$	4.291
Distance from drainage	$(0.55)(0.143) + (0.213)(0.143) + (0.193)(0.143) + (0.045)(1) = 0.181$	4.059

A value of 0.089 21 would be considered to represent acceptable consistency within the PCM.

## 18.5 Multi-criteria combination methods

*Multi-criteria evaluation (MCE)* is a process in which multiple layers are aggregated to yield a single output map or *index of evaluation*. Often this is a map showing the suitability of land for a particular activity. It could be a hazard or prospectivity map or some other parameter that is a function of multiple criteria. Several methods are described here, in order of complexity: Boolean combination, index-overlay, algebraic combination, weights of evidence modelling (Bayesian probability), Dempster–Shafer theory (DST), Weight linear factors in combination (WLC), otherwise known as the analytical hierarchy process (AHP), fuzzy logic and vectorial fuzzy modelling. Weights of evidence modelling, WLC, AHP, vectorial fuzzy modelling and DST can all be considered as providing fuzzy measures since all allow uncertainty to be incorporated in some way, either directly through the use of fuzzy membership sets, or through probability functions or some other gradational quantities.

### 18.5.1 Boolean logical combination

This represents the simplest possible method of factor combination. Each spatial relationship is identified and prepared as a map or image where

every location has two possible conditions: suitable or unsuitable. One or more of the standard arithmetic operators is used to combine the spatial relationship factors into a single map. The Boolean AND combinatorial operator is most commonly used, and retains only those areas that are suitable in all input factors. Alternatively, the combinatorial Boolean OR can also be used, which represents the conceptual opposite of AND, and will always result in more or larger areas being categorized as suitable. The former represents the ‘risk averse’ or conservative of the two methods, and the latter is the more ‘risk taking’ or liberal of the two. This method is simple, conceptually and computationally, but is perhaps rather oversimplistic, since there is no allowance for gradational quantities or for other forms of uncertainty.

### 18.5.2 Index-overlay and algebraic combination

Here criteria are still categorical but they may comprise more than two discrete levels of suitability. These are usually represented as ordinal-scale numbers, so that a location with a value of 2 is more suitable than a location with a value of 1 but is not twice as suitable. The resultant suitability map is constructed by the summation of all the input factors: the higher the number, the more suitable the location. This approach is also simple and effective but also has some drawbacks, since the imposed criteria classes are subjective and they behave like weights so that a factor divided into 10 levels will have greater effect on the result than

one divided into only 3 levels. Ideally, therefore, the input datasets should be scaled to the same number of classes. The result is also unconstrained in that an increase in the number of input criteria causes an increase in the range of values in the suitability map.

The criteria are combined using simple summation, or arithmetic or geometric mean operators, according to the desired level of conservatism in the result. Use of the arithmetic mean is more liberal in that all criteria and locations pass through to the result even if a zero is encountered. The geometric mean can be considered more conservative since any zero value causes that location to be selectively removed from the result. These different operators should be used selectively to combine input criteria in different decision-making situations. For instance, where there is considerable confidence about the particulars of the case, i.e. there is decision rule confidence, a more risk-taking geometric mean method might be applicable. Conversely, if there are plentiful data but a great deal of decision rule uncertainty, the more risk-averse arithmetic mean may be appropriate, allowing all values and positions through to the end result. A research case study, in which an index-overlay combination method based on the geometric mean is used for landslide hazard assessment, is presented in Section 21.2.

The index-overlay method can be modified and improved on by replacing the ordinal-scale numbers with ratio-scale numbers, so that a value of 2 means the location is twice as suitable as a location with a value of 1; this removes the need for further scaling. Criteria combination can then proceed in the same way by summation or arithmetic and geometric mean.

### 18.5.3 Weights of evidence modelling based on bayesian probability theory

One of the most widely used statistical, multi-criteria analysis techniques is the *weights of evidence method*, which is based on *Bayesian probability*. Here the quantitative spatial relationships between datasets representing significant criteria (input evidence) and known occurrences (outcomes) are analysed using Bayesian weights of evidence probability analysis. Predictor maps and layers are used as

input evidence. The products are layers representing the estimated probability of occurrence of a particular phenomenon (according to a hypothesis) and of the uncertainty of the probability estimates. This involves the calculation of the likelihood of specific values occurring or being exceeded, such as the likelihood of a pixel slope angle value exceeding a certain threshold, as part of a slope stability assessment.

Bayesian probability allows us to combine new evidence about a hypothesis on the basis of prior knowledge or evidence. This allows us to evaluate the likelihood of a hypothesis being true using one of more pieces of evidence. Bayes' theorem is given as

$$p(h|e) = \frac{p(e|h) \times p(h)}{\sum_n p(e|h_n) \times p(h_n)} \quad (18.9)$$

where  $h$  represents the hypothesis and there are  $n$  possible, mutually exclusive, statistically independent outcomes;  $e$  is the evidence (some kind of observation or measurement);  $p(h)$ , the *prior probability*, represents the probability of the hypothesis being true regardless of any new evidence; whereas  $p(e|h)$  represents the probability of the evidence occurring given that the hypothesis is true, i.e. the *conditional probability*; and  $p(h|e)$ , the *posterior probability*, is the probability of the hypothesis being true given the evidence.

One assumption considered here is the independence, or lack of, between the input evidence layers. If two evidence layers,  $e_1$  and  $e_2$ , are statistically independent, then the implied probabilities of their presence are

$$p(e_1|e_2) = p(e_1) \text{ and } p(e_2|e_1) = p(e_2); \quad (18.10)$$

i.e., the conditional probability of the presence of  $e_1$  is independent of the presence of  $e_2$ , and vice versa. If however the two variables are *conditionally independent* with respect to a third layer,  $t$ , then the following relationship exists:

$$p(e_1 \cap e_2|t) = p(e_1|t)p(e_2|t). \quad (18.11)$$

If  $e_1$  and  $e_2$  are binary evidence layers for an area and  $t$  represents known target occurrences in that area, then the following allows us to estimate the number of targets that might occur in the area of

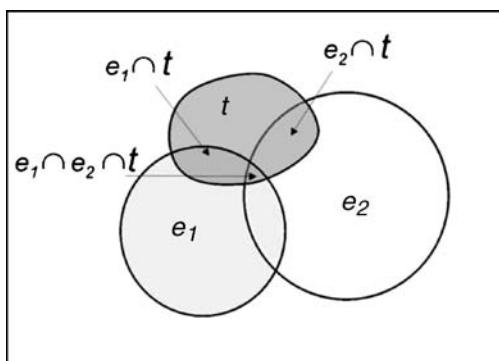
overlap between  $e_1$  and  $e_2$  (i.e. where both are present):

$$n(e_1 \cap e_2 \cap t) = \frac{n(e_1 \cap t)n(e_2 \cap t)}{n(t)} \quad (18.12)$$

where the predicted number of target occurrences in the overlap will equal the number of target occurrences in  $e_1$  times the number of target occurrences in  $e_2$ , divided by the total number of target occurrences, if the two variables are conditionally independent (Figure 18.3). If the total estimated (predicted) number of targets is larger than the actual number of occurrences, then the conditional independence can be considered to be ‘violated’ and the input variables being compared should be checked.

When input layers are being prepared at the start of the MCE process, the binary evidence layers should be compared, for example in a pairwise fashion, to test for conditional independence. If necessary, problematic layers should be combined to reduce the conditional independence effect or they should be removed.

These descriptions seem rather abstract in themselves, so we could consider a simple example, with one target occurrence dataset and one evidence layer, to explain the principle better. If we take mineral exploration as an example, and a dataset of 100 samples, represented as a raster of  $10 \times 10$  pixels; of the 100 samples, there are five gold occurrences. So there are two possible mutually exclusive outcomes, containing gold or not containing gold, i.e.  $n = 2$ .



**Figure 18.3** Schematic illustration of the effect of conditional independence

It appears that the probability of finding gold is 0.05 and that of not finding gold, 0.95; this is the prior probability of finding gold in this area. The new evidence introduced in this case is a layer representing a geophysical parameter (such as total magnetic field). This layer contains an area of anomalously high values so a reclassified evidence layer, containing two classes (anomaly or no anomaly), is introduced. It is found that four of the five gold occurrences lie within the area characterized as anomalous, so it seems that the chances of encountering gold are much higher given the presence of the geophysical anomaly since 0.8 of the pixels that contain gold are also on the anomaly. Of the 100 pixels, 95 contain no gold but 25 of these are also geographically coincident with the anomaly, so that 0.275 of pixels with no gold occur in the anomaly. Using these values, the probability of finding gold given the presence of the anomaly, or the posterior probability, will therefore be

$$\begin{aligned} p(h|e) &= \frac{0.80 \times 0.05}{(0.80 \times 0.05) + (0.275 \times 0.95)} \\ &= \frac{0.04}{0.04 + 0.261} = 0.133 \end{aligned} \quad (18.13)$$

which is greater than 0.05, the  $p(h)$ . And in a similar way, the posterior probability of finding gold where there is no measured anomaly will be

$$\begin{aligned} p(h|e) &= \frac{0.20 \times 0.05}{(0.20 \times 0.05) + (0.725 \times 0.95)} \\ &= \frac{0.01}{0.01 + 0.689} = 0.0143 \end{aligned} \quad (18.14)$$

which is less than 0.05, the  $p(h)$ .

So, after the introduction of new evidence (in this case the presence of the anomaly), the posterior probability of finding gold is considerably greater than without that evidence: 0.133 as opposed to 0.05. Therefore the introduction of various new pieces of evidence (representing significant criteria) to the analysis increases our chances of finding the target, and of making good decisions.

One limitation of this simple approach is that gradational outcomes are not permitted, and only two states of nature can exist: the pixel either contains gold or does not contain gold. The outcomes are said to be mutually exclusive. No account

of the variable amount of gold present can be made. Similarly, the evidence layer contains only two classes, anomalous or not anomalous, meaning that a subjective judgement has been made in the reclassification of this input evidence. To incorporate gradational values we consider each input as a continuous normal distribution function rather than one with discrete levels. Probabilities are then derived as multivariate vector and matrix calculations to give gradational predictions.

A method of factor combination based on Bayesian probability is described in Section 22.1. This case, set in south-east Greenland, involves prospectivity mapping for a number of commodities, in which there are reliable but limited data available and reasonably good understanding about the deposit models.

#### 18.5.4 Belief and Dempster–Shafer theory

DST is a knowledge-driven approach based on *belief functions* and *plausible reasoning*, and is used to combine separate pieces of evidence to calculate the probability of an event or occurrence. The theory owes its name to work by Shafer (Shafer, 1976) in extending the Bayesian theory of statistical inference (Dempster, 1967 & 1968). The mathematical derivations are also dealt with in great detail in a number of other texts.

A limitation of many of the approaches described so far is the assumption (requirement) that all the input factors must contribute positively towards the outcome; high input values are correlated with suitability. As a consequence, there is no way to consider evidence that seems to be contradictory to the main hypothesis. DST allows the estimation of the likelihood of suitability, or unsuitability, in addition to estimates of plausibility and belief for any hypothesis being evaluated. In traditional Bayesian probability the absence of supporting evidence for a hypothesis is automatically assumed to support the alternative hypothesis, unsuitability. The DST is unique in allowing that ‘ignorance’ does not necessarily support that alternative hypothesis and in doing so provides a method for identifying areas and ways to reduce uncertainty.

DST introduces six quantities: basic probability assignment, ignorance, belief, disbelief, plausibil-

ity and a belief interval. It also provides estimates of the confidence levels of the probabilities assigned to the various outcomes. The degree to which evidence supports a hypothesis is known as *belief*, the degree to which the evidence does not contradict that hypothesis is known as *plausibility*, and the difference between them is referred to as the *belief interval*; the last serves as a measure of uncertainty about a particular hypothesis. Belief is always less than or equal to plausibility. Belief in a hypothesis is the sum of the probability ‘masses’ of all subsets of the hypothesis. Plausibility is therefore an upper limit on the possibility that the hypothesis could happen, i.e. it ‘could possibly happen’ up to that value, because there is only some evidence that contradicts this hypothesis. Plausibility represents 1 minus the sum of the probability masses of all sets whose intersection with the hypothesis is empty or, in other words, the sum of the masses of all sets whose intersection with the hypothesis is *not* empty. A degree of belief or *mass* is represented as a belief function rather than a probability distribution. Probabilities are therefore assigned to sets of possibilities rather than to single, definitive occurrences and this incorporates a measure of uncertainty.

Using another very simple example (see Table 18.6), considering the hypothesis that a pixel position contains no gold, we may have evidence suggesting that the pixel area contains no gold, with a confidence of 0.5, but the evidence contrary to that hypothesis (i.e. pixel contains gold) only has a confidence of 0.2. So for our hypothesis we have a belief of 0.5 (lower limit) and a plausibility of 0.8 (upper limit). The remaining mass of 0.3 (the gap between the 0.5 and 0.2) represents the probability that the pixel may or may not contain gold; this interval represents the level of uncertainty caused by a lack of the evidence for the hypothesis.

The null hypothesis is zero by definition and this represents ‘no solution’ to the problem. The mutually exclusive hypotheses ‘Gold’ and ‘No gold’ have probabilities of 0.2 and 0.5, respectively. The universal hypothesis ‘Either’ represents the assumption that the pixel contains something (gold or not) and forms the remainder so that the sum of the probability masses is 1. The belief value for ‘Either’ consists of the sum of all three probability masses (‘Either’, ‘Gold’ and ‘No gold’)

because ‘Gold’ and ‘No gold’ are subsets of ‘Either’, whereas the belief values for the ‘Gold’ and ‘No gold’ hypotheses are equal to their individual probability masses since they have no subsets. So the plausibility of ‘Gold’ occurring is equal to the sum of probability masses for ‘Gold’ and ‘Either’; the ‘No gold’ plausibility is equal to the sum of probability masses of ‘No gold’ and ‘Either’; and the ‘Either’ plausibility is equal to the sum of probability masses of ‘Gold’, ‘No gold’ and ‘Either’. The hypothesis ‘Either’ must always have 100% belief and plausibility and so acts as a ‘checksum’ on the result.

Evidence layers are then brought together using *Dempster’s rule of combination*, which is a generalization of Bayes’ theorem. The combination involves the summation of two input probability masses and normalization to 1. If we use a simple example with two input datasets ( $i$  and  $j$ ), then

$$e_c = \frac{(e_i e_j) + (e_i u_j) + (e_j u_i)}{\beta} \quad (18.15)$$

$$d_c = \frac{(d_i d_j) + (d_i u_j) + (d_j u_i)}{\beta} \quad (18.16)$$

$$u_c = \frac{u_i u_j}{\beta} \quad (18.17)$$

where  $e$  = evidence (or belief) for input evidence datasets  $i$  and  $j$ ,  $d$  = disbelief,  $u$  = uncertainty (which equals plausibility minus belief),  $e_c$  = the combined evidence,  $d_c$  = the combined disbelief,  $u_c$  = the combined uncertainty and  $\beta$  = the normalization factor. Here  $\beta$  is derived as

$$\beta = 1 - e_i d_j - d_i e_j. \quad (18.18)$$

This method tends to emphasize the agreement between input evidence and ignore all conflicts between them via the normalization factor. The latter ensures that evidence (belief), disbelief and uncertainty always equal 1 ( $e + d + u = 1$ ). Where conflicts between input evidence are known to be of significant magnitude, DST can produce meaningless results and an alternative method should be sought.

### 18.5.5 Weighted factors in linear combination

This method is sometimes referred to as the *analytical hierarchy process* (Saaty, 1990), a decision-making technique that allows consideration of both qualitative and quantitative aspects of decisions. The AHP method rather importantly accepts that:

- Certain criteria are more important than others.
- Criteria have intermediate values of suitability (i.e. they do not need to be simply classed as ‘suitable’ or ‘unsuitable’).

Criteria can also be coded to behave differently. Those acting as Boolean ‘constraints’ can be used selectively to remove or ‘zero-out’ locations and regions. Others, which are continuous or variable in nature, are referred to as ‘factors’. The factors should be combined in a way that reflects the two points above and there are three important issues surrounding this combination: determination of the relative importance of each criterion; the standardization or normalization of each factor (since each must be on a consistent scale and must contribute in the same direction towards suitability); and the method of factor combination.

It is assumed to be unlikely that all factors will have equal effect on the outcome. The assessment of their relative importance involves derivation of a series of weighting coefficients. These factor weights control the effect that each factor has on the outcome. The relative significance of each factor, its influence on the other factors and on the outcome need to be compared. This involves ordering the factors into a hierarchy of significance, and assessment of their degree of influence on the outcome and on each other.

The WLC method allows a measure of suitability in one factor to be compensated for in another factor(s) through the use of weights. The general approach of this method is as follows:

1. Identify the criteria (decide which criteria are factors and which are constraints).
2. Produce an image or coverage representing each criteria.
3. Standardize or scale each factor image (for instance, using fuzzy functions), reclassifying to a real-number scale (0 to 1), byte scale (0 to 255) or

percentage scale (0–100). All must be scaled to the same range and in the same direction so that they each contribute either positively or negatively (generally the former) towards the outcome.

4. Derive weighting coefficients which describe the relative importance of each factor (by one of the methods already described, usually by the PCM method).
5. Linearly combine the factor weights with the standardized factors and the constraints (usually by aggregation) to produce the ‘suitability’ map.

The consequence of using an aggregation method is that all candidate pixels entering the model pass through to the end. At no point does an encountered zero value cause termination of the model for that pixel position. So while the factors are weighted in terms of their significance, there is a total ‘trade-off’ between each factor value encountered: that is, a low score in one factor can be compensated for by a very high score in another.

The WLC method is popular but can in some instances be considered too ‘liberal’ in its handling of the data in the system, since it involves equal ranking of the weighted factors and allows full trade-off between them. Its factor aggregation method can be likened to a parallel connection system, which allows all input criteria to survive to the end (the likelihood of the occurrence of a zero is low). It is also possible that the relationships between the input factors are not linear, in which case a more complex model will be required. In many cases, this parallel system may be appropriate but in others a harsher, more risk-averse system may be better, one which enables certain factors or combinations of factors to be eliminated

completely from the system, rather like a sequential connection system. Factor combination via calculation of the geometric mean (as opposed to the arithmetic mean) represents such a system in which the occurrence of a zero rating terminates the system and eliminates that location from the analysis.

An illustration of this method applied using fuzzy scaled inputs (see also Section 15.5.6) is described in a research case study, in Section 21.3. The index-overlay combination introduced in Section 18.5.2 is the very simplest case of WLC.

#### 18.5.5.1 Ordered weighted average

The ordered weighted average (OWA) represents a refinement of the WLC method, where the degree of trade-off between factors is controlled by a second set of *order weights*. With full control over the size and distribution of the order weights, the amount of risk taking and degree of trade-off (or *substitutability*) can be varied. Trade-off represents the degree to which a low score in one criterion can be compensated for by a higher score in another. The order weights define the rank ordering of factors for any pixel; they are not combined in the same sequence everywhere.

In this way, the degree to which factors can pass through the system is also controllable. Using one arrangement of order weights, pixels may be eliminated in some areas but permitted through in others. Using another arrangement, all pixels may be permitted through (equivalent to WLC).

After the first set of factor weights has been applied, the results are ranked from low to high (in terms of their calculated ‘suitability’ value). The factor with the lowest suitability score is then assigned the first order weight and so on up to the

**Table 18.6** Probability masses associated with the hypothesis that a pixel position contains no gold, under Dempster–Shafer theory

Hypothesis	Probability	Belief	Plausibility
Null (no solution)	0	0	0
Gold	0.2	0.2	0.5
No gold	0.5	0.5	0.8
Either (gold or no gold)	0.3	1.0	1.0

factor with the highest suitability being assigned the highest order weight. The relative skew to either end of the order weights determines the level of risk, and the degree to which the order weights are evenly distributed across the factor positions determines the amount of trade-off (see Table 18.7).

For example, consider three factors  $a$ ,  $b$  and  $c$ , to which we apply weights of 0.6, 0.3 and 0.1, on the basis of their rank order (the order weights sum to 1.0). At one location the factors are ranked  $cba$ , from lowest to highest, and the weighted combination will be  $0.6c + 0.3b + 0.1a$ . If at another location, however, the factors are ranked  $bac$ , the weighted combination will be  $0.6b + 0.3a + 0.1c$ . A low score in one factor can therefore be compensated for by a high score in another; there is a trade-off between factors.

Two parameters, *AND/ORness* and *TRADEOFF*, are used to characterize the nature of an OWA operation:

$$\text{TRADEOFF} = 1 - \sqrt{\frac{n \sum (w_i - 1/n)^2}{n - 1}} \quad (18.19)$$

$$\text{ANDness} = (1/(n-1)) \sum ((n-i)w_i) \quad (18.20)$$

$$\text{ORness} = 1 - \text{ANDness} \quad (18.21)$$

where  $n$  is the total number of factors,  $i$  is the order of factors and  $w_i$  is the weight for the factor of the  $i$ th order. From the equation, *ANDness* or *ORness* is governed by the amount of skew in the order weights and the trade-off is controlled by the degree of dispersion in the order weights.

For a risk-averse or conservative result, greater order weight is assigned to the factors nearest the minimum value. For a risk-taking or liberal result, full weighting is given to the maximum suitability

score. If full weight is given to the factor with minimum suitability score and zero to all other positions, then the result will resemble that produced by the Boolean Min(AND) combination of factors and will represent no trade-off between factors. If full weight is given to the maximum suitability score, then the result resembles Boolean Max(OR). If all order weights are equal fractions of 1, then full trade-off is allowed and the result is equivalent to the WLC.

Order weights control the position of the aggregation operator on a continuum between the extremes of Min and Max, as well as the degree of trade-off. Examples are shown in Table 18.7 and illustrated conceptually by the decision strategy space shown in Figure 18.4.

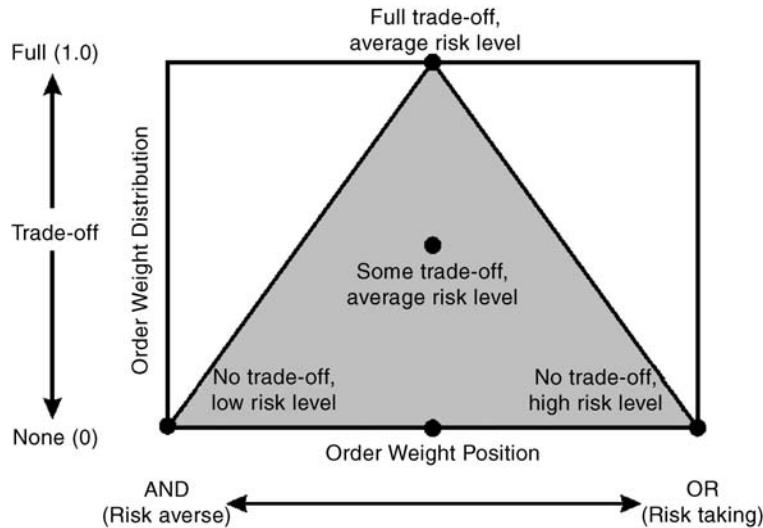
### 18.5.6 Fuzzy logic

Although the Bayesian and algebraic methods can be adapted to accommodate the combination of spatial continuously sampled data, there is a branch of mathematics, *fuzzy logic*, which is ideally suited for this purpose. Fuzzy logic represents a ‘superset’ of Boolean logic, and deals with variables that incorporate some uncertainty or ‘fuzziness’. Since fuzzy membership removes the requirements of ‘total membership’ of a particular class, it provides an ideal way to allow for the *possibility* of a variable being suitable or unsuitable. We have already described how this can be used to incorporate uncertainty within our analysis in the previous chapter.

The threshold values that define the fuzzy set become the input parameters when preparing each variable in our spatial analysis. The values are chosen based on prior knowledge and understanding of the data and decision rules. The type of

**Table 18.7** Order weights for a five-factor example used by OWA

Rank	1st	2nd	3rd	4th	5th	Description
Order weight	1	0	0	0	0	Risk averse, no trade-off
Order weight	0	0	0	0	1	Risk taking, no trade-off
Order weight	0	0	1	0	0	Average risk, no trade-off
Order weight	0.2	0.2	0.2	0.2	0.2	Average risk, full trade-off



**Figure 18.4** Decision strategy space in order weighted averaging. Modified after Saaty (1990)

membership function we choose will depend on the way the phenomenon contributes to the outcome; positively or negatively, monotonically or symmetrically. The threshold values are applied according to the user's understanding and 'ground' knowledge of the phenomenon, to best reflect their significance on the result. An underlying assumption at this point is that the relationship between the input factors and the outcome is linear, but this may not necessarily be so.

Each input factor is scaled according to the chosen fuzzy membership function in preparation for factor combination. All the factors must be scaled in the same direction, i.e. they must all contribute to the outcome in the same way, either positively or negatively. In this way, the locations which represent the most desirable characteristics are all coded with either very high values or very low values according to choice. An illustration of the WLC method applied using fuzzy scaled inputs is described in a research case study in Section 21.3.

The resultant fuzzy factor layers are then combined in one of a number of ways. The simplest option is via simple set theory using map algebra, such as using Boolean logical intersection (AND) and logical union (OR) operators.

A series of fuzzy operators have subsequently been developed around set theory, for the combination of scaled ('fuzzified') input factors, namely

fuzzy AND, fuzzy OR, fuzzy algebraic product, fuzzy algebraic sum and a fuzzy gamma function:

$$\text{Fuzzy AND} \quad \mu_c = \min(\mu_1, \mu_2, \mu_3, \dots, \mu_n) \quad (18.22)$$

$$\text{Fuzzy OR} \quad \mu_c = \max(\mu_1, \mu_2, \mu_3, \dots, \mu_n) \quad (18.23)$$

$$\text{Fuzzy NOT} \quad \bar{\mu} = 1 - \mu \quad (18.24)$$

where  $\mu_c$  represents the combined fuzzy membership function of  $n$  individual fuzzy inputs. These operate in much the same manner as the Boolean versions. The AND operator produces the most conservative result, producing low values, and allows only areas which are characterized by favourable conditions in all input layers to survive to the end result. In contrast, the OR operator produces the most liberal or risk-taking result and is suitable when it is desirable to allow any favourable evidence to survive and be reflected in the end result.

This simple combination may not be considered suitable for the combination of multiple datasets because it is possible that extremely high and low values can propagate through to the final result. Two operators have been developed to overcome this problem: the *fuzzy algebraic product* (FAP) and the *fuzzy algebraic sum* (FAS). The FAP is the combined

product of all the input values or fuzzy factors in the following way:

$$\mu_c = \prod_{i=1}^n \mu_i \quad (18.25)$$

where  $\mu_c$  represents the FAP fuzzy membership function for the  $n$ th input factor. Since the values being combined are all fractions of 1, the values in the final result are always smaller than the lowest contributing value in any layer. The function can be considered ‘decreasive’ for this reason. The FAS is not a true sum and is derived by

$$\mu_c = 1 - \prod_{i=1}^n (1 - \mu_i). \quad (18.26)$$

Here the reverse is true: the resulting value will always be larger than the largest contributing value in any layer, but is limited by the maximum value of 1. This function therefore has the opposite effect to the FAP and is considered ‘increasive’. Two pieces of input evidence which favour the result would reinforce one another in this method. It is worth noting here that the output value is partly affected by the number of input datasets: the more the number of datasets, the greater the resulting value. FAP and FAS can also be combined into a single operation, called a *gamma function*, which is calculated as

$$\begin{aligned} \mu_c &= [\text{FAS}]^\gamma [\text{FAP}]^{1-\gamma} \\ &= \left[ 1 - \prod_{i=1}^n (1 - \mu_i) \right]^\gamma \left[ \prod_{i=1}^n \mu_i \right]^{1-\gamma} \end{aligned} \quad (18.27)$$

where the gamma parameter ( $\gamma$ ) varies between 0 and 1. When a value of 0 is chosen, the result is equivalent to the FAP. When gamma is 1.0 the result is equivalent to the FAS. A gamma value somewhere between provides a compromise between increasive and decreasive tendencies of the two separate functions. This method allows for uncertainty to be incorporated and allows all the input factors to contribute to the final result but has the drawback that all the input factors are treated equally. They must also contribute in the same direction towards the outcome.

In this way, pieces of evidence can be combined sequentially, in a series of carefully designed steps, rather than in one simultaneous operation. This

gives more control over the final outcome and allows the different input layers to be treated differently, according to the understanding of the layer’s contribution to the outcome.

### 18.5.7 Vectorial fuzzy modelling

In an attempt to improve on the above, the *vectorial fuzzy logic* method has been developed (Knox-Robinson, 2000) in the mapping of mineral prospectivity or suitability. The fuzzy vector is defined by two values, the calculated prospectivity (the fuzzy vector angle) and confidence (the fuzzy vector magnitude), the latter as a measure of similarity between input factors. Using the vectorial fuzzy logic method, null data and incomplete knowledge can be incorporated into the multi-criterion analysis. The ‘confidence’ value actually performs several functions: it represents confidence in the suitability value; the importance of each factor relative to others; and it allows null values to be used. The combination of the two values involves calculating a vector for each spatial relationship factor. The combined lengths and directions of each vector provide the aggregate suitability. The closer the value of the inputs, the longer the resultant combined vector ( $c_c$ ) and the higher the confidence level. Confidence is a relative measure of consistency throughout the multi-criteria dataset for any particular location. The two values are derived as follows:

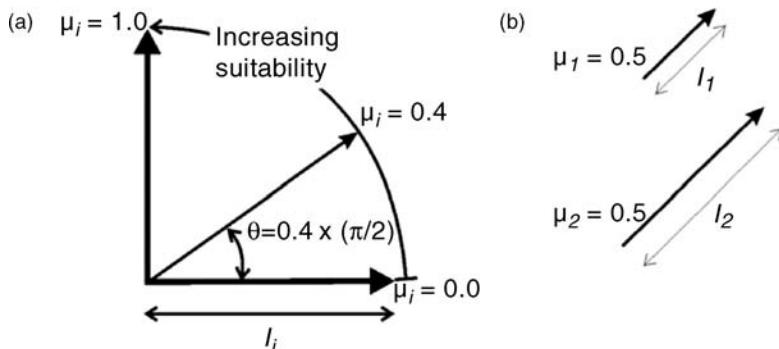
Fuzzy prospectivity :

$$\mu_c = \left( \frac{2}{\pi} \right) \arctan \left[ \frac{\sum_{i=1}^n c_i \sin(\pi \mu_i / 2)}{\sum_{i=1}^n c_i \cos(\pi \mu_i / 2)} \right] \quad (18.28)$$

Fuzzy confidence :

$$c_c = \sqrt{\left[ \sum_{i=1}^n c_i \sin\left(\frac{\pi \mu_i}{2}\right) \right]^2 + \left[ \sum_{i=1}^n c_i \cos\left(\frac{\pi \mu_i}{2}\right) \right]^2} \quad (18.29)$$

where  $\mu_i$  is the fuzzy suitability value for the  $i$ th factor input layer ( $0 \leq \mu_i \leq 1$ ). Figure 18.5 shows the concept of variable suitability (prospectivity)



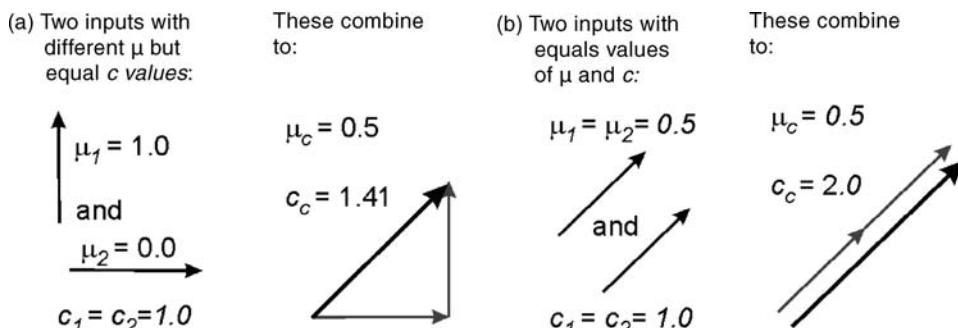
**Figure 18.5** Illustration of variable prospectivity derived using the vectorial fuzzy modelling method. After Knox-Robinson (2000)

represented as a vector using this method. In Figure 18.5 a the vector quantity of suitability ( $\mu$ ) is given as the direction of the vector and the confidence level ( $c$ ) of that suitability value is represented by the vector's length. In Figure 18.5b two examples illustrate vectors of equal suitability (constant direction), but of differing confidence levels (different lengths). The vector ( $\mu_2$ ) is longer and therefore represents a greater level of confidence in the suitability value, and so the corresponding pixel would have a greater influence on the final output value.

In Figure 18.6a, two input criteria with different suitability values but equal confidence levels provide a combined suitability value of 0.5 (the average  $\mu$  value) and a confidence level of 1.41 (derived simply by Pythagorean geometry);

the confidence level is lower since the two input suitability scores are conflicting. In Figure 18.6b the two inputs have identical suitability and confidence level values, so they combine to give the same suitability score of 0.5 but with double the confidence level,  $c_c = 2.0$ ; from this result we can be more confident that the result is more representative of the true suitability at that position.

In this method all input factors must contribute positively towards the outcome. The existence or possibility of a trade-off between input factors is considered undesirable or irrelevant in this method. Its use is therefore most appropriate when there is considerable understanding about the influence of each piece of evidence in leading to the outcome; i.e., where there is considerable



**Figure 18.6** Examples of the combination of two 'fuzzy' vectors: (a) two input criteria fuzzy vectors with equal confidence but different suitability (prospectivity); and (b) two inputs with identical suitability and confidence levels. Modified after Knox-Robinson (2000)

decision rule confidence but perhaps some data uncertainties.

## 18.6 Summary

We have summarized in this chapter some of the better known methods of criteria combination; the list is not exhaustive, there are many variations and, regardless of the level of complexity involved, they all have strengths and weaknesses. It may be desirable, and is often appropriate, to use more than one of the criteria combination methods described here, i.e. to use a mixture of fuzzy and non-fuzzy operators within the same model. Certainly, none of these methods can be considered correct or incorrect when it comes to integrating spatial datasets or combining multiple criteria. The important issue is that each method is appropriate to a particular set of circumstances and objectives, which should be carefully considered before choosing that method.

It is important to note, however, that multi-criteria analysis is often carried out with little consideration of the meaning, relevance or quality of the output solution or the effect of potential errors it contains. Some methods for cross-validation, error propagation and sensitivity analysis have already been described (in Chapters 15 and 17) and they are especially relevant here too. The object of spatial analysis in these circumstances is to predict conditions beyond the location or times where information is available. The multi-criteria evaluation procedures discussed

here provide a means to allow for, quantify and reduce certain varieties of uncertainty in achieving these predictions, but not all types. So it is in the final stages of such analysis, more than any other, that the validity of the ‘prediction’ should be questioned and tested. The paradox is that there is always a solution but never a perfect solution and we always want to know exactly how good a solution is while the answer will for ever be fuzzy.

## Questions

- 18.1 What are the assumptions made in choosing a weighting method?
- 18.2 Why is it important to know the measurement scale of the input criteria?
- 18.3 What is the difference between order weights and factor weights?
- 18.4 Which weighting method is the most appropriate for a group decision-making process?
- 18.5 Why are certain weighting methods subject to uncertainty?
- 18.6 What are the chief differences between a simple Boolean model and one that incorporates uncertainty?
- 18.7 What is the difference between probability and possibility, and how do these two concepts help us in spatial analysis?
- 18.8 What more could be done to improve the model and/or the validity of the result?

# Part Three

## Remote Sensing Applications

In Parts One and Two of this book, we learnt the essential image processing and GIS techniques. Here we will demonstrate, step by step, with examples, how these techniques can be used effectively in remote sensing applications. Although many case studies are drawn from our own research projects in earth sciences and the terrestrial environment, it is important to stress the generic sense of these examples in terms of concept and methodology for wider applications. From this viewpoint, our aim is not to provide rigid recipes for fixed problems but to provide guidance on ‘how to think’ and ‘how to approach’. When first presented with a new project, beginners and students may feel a little lost, rather as in the Chinese saying ‘the tiger wants to eat the sky but doesn’t know where to bite’. We will be satisfied if this part serves as a catalyst, to get you on the track of a real remote sensing application project.



# 19

## Image Processing and GIS Operation Strategy

In this chapter, we describe how the processing, interpretation and analysis of multi-source image and map data (in general, raster datasets) should be approached to produce thematic maps for a typical project. Following the discussion of basic strategy, a simple example of digital geological mapping, based on processing and interpretation of Landsat-7 ETM+ imagery, is presented to demonstrate the work flow from image processing to map composition.

We suggest the following rules of thumb as general guidance for operational strategy:

- *Purpose:* The aims and objectives of the project should be the driving force behind the image processing and multi-source data manipulation. In other words, it should be application driven rather than data processing driven. This is different from algorithm development, which may be triggered by application requirements but is focused on the technical part of data processing and its effectiveness; application examples serve as demonstrators of the algorithm.
- *Keep things simple:* Recall from Figure P.1 in Part One that this is not only true for image processing but serves as good advice for an application project. Nowadays image processing and GIS software packages are so functional and

are supported by ever-increasing computing power. It is far too easy to be dragged into a complicated ‘computer game’ rather than focus on the central theme of the project and to produce the required result in the simplest and most effective way. For learning purposes, we encourage students to experiment with all the relevant processing techniques, while in a real operational case the simplest, cheapest and most effective method is the best choice. Simplicity is beauty.

- *From simple to complex:* Keeping things simple does not necessarily mean they can be achieved in a simple way. If simple image processing and GIS techniques were adequate for all applications, more complicated and advanced techniques would never be developed! Some key information can only be enhanced, analysed and abstracted using complex algorithms and methodology. Starting at the simple end, with general image enhancement techniques, will allow better comprehension of the scope and true nature of the task. Only then should complex techniques be configured and employed to reveal the specific, diagnostic features of the intended targets, and to extract the most crucial information.
- *Reality checks:* After you have performed some processing and produced an exciting-looking result, you should always ask yourself if the result

is realistic: does it make sense? Performing this kind of ‘reality check’ could involve correlating the result with simpler images (e.g. colour composites), other forms of analysis and/or published information (if available). Such information may in itself be insufficient, out of date, at too gross a scale or geographically incomplete but, when compared with your result, may collectively point to your having produced something useful and realistic. If all scant pieces of evidence point in the same direction, you should be on the right track!

- *Relationship between analysis and visualization:* A remote sensing application project normally begins with image visualization and its final results are often in the form of maps and images, for which visualization is again necessary. As a 2D array of numbers, a digital image can easily be numerically analysed. The results of the analysis are not necessarily raster datasets but they can always be visualized in one way or another. In general, we are far more able to comprehend complicated information graphically than numerically. For remote sensing applications, visualizing and interpreting the results of every stage of image processing and GIS analysis are essential to help assess your progress and to decide on the next step towards the final goal of the project.
- *Thinking in three dimensions:* Remote sensing deals with nothing but images, and an image is a 2D representation of 3D phenomena. For centuries, we have tried every possible approach to presenting the 3D Earth on a 2D plane, that is projected maps, and therein lies the science and engineering of geomatics and geodesy. Now, thanks to the development of computer graphics, moving between 2D and 3D representations is much easier. The digital form of topographic maps, namely the DEM, is in itself a 2D raster dataset representing the 3D topography of the Earth’s surface. Using the powerful 3D graphical functions in modern image processing and GIS software packages, we can easily simulate the 3D environment using 2D data, by draping multi-spectral satellite colour composite images over a 3D perspective view of a DEM (as described in Chapter 16). Thinking and viewing in three dimensions can make some tricky information in 2D images suddenly obvious and clearly understandable. In fact it may reveal information that was completely unknown from 2D

observation. For instance, a low-angle reverse fault appears to be a curved line in an image, depending on its intersection relationship with topographic slopes. The 3D thinking and visualization make you realize that it is a low-angle planar surface rather than a steeply dipping curved one.

## 19.1 General image processing strategy

Image processing is almost always the first step of any remote sensing application project but it is often given greater significance than it deserves. In fact, one of the main objectives of image processing is to optimize visualization of a particular thematic dataset. Visual interpretation is therefore essential. Thematic maps are the most important products of remotely sensed imagery, and they are derived by either visual interpretation or image segmentation (computerized classification). Thus far, broadband multi-spectral and SAR images are the most commonly used datasets. The image processing strategy proposed in this section is most relevant to these types of data, and its goal is the effective discrimination of different spectral and spatial targets. We use the word ‘discrimination’ advisedly in this context; in general, it is only possible to differentiate between rocks, soils and mineral groups using broadband image data, rather than identify them.

In contrast, the processing of hyperspectral image data is to achieve spectral target identification, to species level in the case of rock-forming minerals, and thus has a different strategy. Many people make the mistake either of thinking that hyper-resolution is the answer to all problems, or of being put off investing in such technology at all because they do not understand its role or are suspicious of its acclaimed capability. A hyperspectral dataset is acquired using an imaging spectrometer or hyperspectral sensor, which is a remote sensing instrument that combines the spatial presentation of an imaging sensor with the analytical capabilities of a spectrometer. Such a sensor system may have up to several hundred narrow bands, with a spectral resolution of the order of 10 nm or narrower. Imaging spectrometers produce a near-complete spectrum for every pixel of the image, thus allowing the specific identification of materials rather than merely the discrimination between them. A hyperspectral

dataset truly provides a data volume or cube. Here it is more important to analyse the spectral signature of each pixel than to perform general image enhancement. The processing methodology and strategy are therefore very different from broadband image processing in many aspects, although the enhancement for image visualization is still important. Considering that hyperspectral remote sensing is a broad and important topic on its own, covering data processing and application development, in this book we have decided to discuss it only briefly and to focus instead on broadband multi-spectral remote sensing.

When you begin a project, you should think along the following lines and, broadly speaking, in the following order:

1. What is the application theme and overall objective of the project?
2. What kind of thematic information do I need to extract from remotely sensed images?
3. At what scale do I need to work? In other words, what is the geographic extent and what level of spatial or spectral detail is required within that area?
4. What types of image data are required and available?
5. What is my approach and methodology for image/GIS processing and analysis?
6. How do I present my results (interpretation and map composition)?
7. Who will need to use and understand the results (technical, managerial or layperson)?

Once these steps have been thought through and the data have been acquired, the generation of thematic maps from remote sensing image data is generally carried out in three stages:

1. Data preparation.
2. Processing for general visualization and thematic information extraction.
3. Analysis, interpretation and finally map composition.

In the following sections, we describe the thematic mapping procedure in a linear sequence for clarity. In reality, the image processing and image interpretation are dynamically integrated. The interpretation of the results of one stage of image processing may lead to the image processing and

data analysis of the next stage. Often, you may feel you have reached the end, after producing a wonderful image; the subsequent interpretation of that image may then spur you on to explore something further or something completely different. A thematic map derived from remotely sensed image may be used alone or as an input layer for further GIS analysis, as outlined in Section 19.2.

### 19.1.1 Preparation of basic working dataset

#### 19.1.1.1 Data collection

At the stage of sensor development, sensor configuration (spectral bands, spatial resolution, radiometric quantization, etc.) is decided based on wide consultation of application sectors, in an attempt to provide capabilities that meet actual requirements, subject to the readiness of the technology of course. The data collection is, however, often dictated by what is available, or what the budget will allow, rather than what is actually required. As a result, remotely sensed image datasets are generally aimed at a broad range of application fields and so may not be able to satisfy the most specific needs of some cases. Within the context of these constraints we should therefore always ask the following questions: What is the purpose of the job? And which dataset will be the most cost effective or will provide the most information relevant to the purpose of the job? In many ways, image processing aims to enhance and extract thematic information that is not directly sensed or distinctively recorded in any one single image. Sometimes, it is a matter of detective work!

At these early stages of choosing and preparing the dataset, it is also important to consider the most appropriate mapping scales for particular datasets, or rather to choose the most appropriate data to suite the mapping requirements of the task being undertaken. If working at a country-wide or regional scale, it would be rather unwise to select data of very high resolution (VHR) for the work since the costs of doing so might be prohibitive and would generate huge data volumes that might be unworkable or provide a level of detail that is just unnecessary at that stage. On the other hand, if the first regional-scale work is likely to be followed by a more detailed second phase, of the same geographic extent, it may then be necessary to acquire VHR data for the entire area at the start. Unmanned aerial vehicles, carrying

**Table 19.1** Remotely sensed datasets and their appropriate mapping scales

Dataset	Spatial resolution	Swath	Mapping scales
Airborne optical	<10–50 cm	Variable	<1 : 30 000
Airborne hyperspectral	2–10 m	Variable	1 : 10 000
VHR satellite	0.6–5 m	11–60 km	1 : 5000–1 : 10 000
SPOT-1–4	10–20 m	60 km	1 : 25 000
ALOS AVNIR-2	10 m	70 km	1 : 25 000
ASTER	15, 30 and 90 m	60 km	1 : 30 000–1 : 50 000 +
Landsat TM/ETM	30 m (15 m pan)	185 km	1 : 50 000 (1 : 30 000 with pan) +

VHR satellite = IRS Pan, ALOS PRISM, Ikonos, Orbview-3, Quickbird and WorldView.

VHR sensors, and capable of making vast sorties, are increasingly being used for such country-wide mapping projects. This is, however, not the normal way of doing things for the majority of application cases. Commonly, the regional-scale work is followed by more detailed studies in selected areas, in which case you would then acquire higher resolution imagery for those selected areas. The appropriate mapping scale

of data is dictated largely by the spatial resolution and partly by the swath width or footprint of the dataset. A summary of common remotely sensed datasets (spaceborne and airborne, and medium to very high resolution), and the mapping scales at which they are commonly used, is shown in Table 19.1. Table 19.2 presents a summary of the currently available remote sensing sensors and the corresponding application

**Table 19.2** Present remote sensing systems and their application areas

Dataset \ Application	ENV	AGRI	FOR	GEO	MAR	RISK	PLAN	DEF	UTIL
MERIS	Excellent								
Landsat TM & ETM	Excellent								
ASTER	Excellent								
ALOS AVNIR-2				Excellent					
SAR	Excellent	Useful	Useful	Excellent		Useful		Excellent	
Airborne Radar	Excellent	Useful	Useful	Excellent		Useful		Excellent	
SPOT 1–4	Useful	Useful	Useful	Excellent	Useful	Useful	Excellent	Excellent	
VHR Satellites	Useful	Useful	Useful	Useful	Useful	Useful	Useful	Useful	Useful

ENV – environment; AGRI – agriculture; FOR – forestry; GEO – geology and exploration; MAR – marine and coastal; RISK – hazards and risk; PLAN – cartography and urban planning; DEF - defence, security & infrastructure; and UTIL – utilities, telecoms, media and consumer.

SAR = ERS, Envisat, ALOS PALSAR, Radarsat  
VHR Satellite = IRS Pan, ALOS PRISM, Ikonos, Orbview-3, Quickbird and Worldview-1,

Excellent		Useful in some circumstances	
Useful			

**Table 19.3** Recent and forthcoming launches of remote sensing satellite sensors

Launch year	Satellite sensor	Organization
2006	EROS-B Kompsat-2	ImageSat International Korea Aerospace Research Institute
2007	TerraSAR-X WorldView-1 RADARSAT-2	TerraSAR Digital Globe RadarSat International, MDA corporation and RadarSat2 Info
2008	WorldView-2 GeoEye-1 GOES-O and GOES-P	Digital Globe GeoEye GOES at Boeing
2009	EROS-C Nigeriasat-2 Cryosat	ImageSat International SSTL ESA
2010	TerraSAR-2	TerraSAR

areas they are used for. Table 19.3 presents a selection of current and future remote sensing satellites/sensors expected in the coming few years.

#### 19.1.1.2 Georectification, image co-registration and mosaicing

It is essential to ensure that all datasets being used are georectified or georeferenced. These days all digital Earth observation data are provided as georectified products, which are accurate in  $x$  and  $y$  to some specified degree. They are always supplied in one of a few standard formats, and always conforming to WGS84 data and UTM projection, since these are global standards. If, for some reason, the data are not georectified, then the rectification process will normally be the first image processing step carried out. Most image processing and GIS software now provide a ‘projection-on-the-fly’ facility, which removes the requirement that all data conform to the same specific coordinate system. Provided that the coordinate systems of each input dataset are defined numerically, one dataset can be re-projected automatically to the coordinate system of another dataset and so brought into alignment, visualized together and, if necessary, assembled into an image mosaic.

The geometric accuracy of these Earth observation datasets, as they are delivered, is generally adequate for many applications but quite often a higher degree of accuracy is needed. This need will depend on the application in question and on the spatial resolution of the dataset. For instance, Landsat data can be considered ‘medium-resolution’ data; it enables mapping at about 1 : 40 000 (at best) and its geopositioning is generally accurate to about 50 m on the ground. In contrast, VHR image data (Quickbird, Ikonos and SPOT-5) may enable mapping at better than 1 : 10 000 scale on the basis of spatial resolution. When delivered as raw products these claim a positional accuracy of about 15 m, but to map realistically at such scales, and to do so accurately, the data require not only improved georectification but orthorectification to correct for the image distortions imposed by terrain relief.

In such cases, a user may perform the georectification based on measured positions (ground control points or GCPs) acquired using high-quality GPS. If the data also require orthorectification, then a DEM of suitable quality is also required. The quality of the GPS data collection, the users’ capability and how they document their surveys, rather than the instrument’s capability, are of paramount importance.

Resampling of the image data is inevitable during the georectification process. There is an argument for leaving the georectification as the final step after thematic mapping in order to minimize the errors and distortions introduced by the georectification process. This approach is typically based on an image-processing-focused mentality, rather than on a practical one, and so it rather depends on what you intend to do with the data, as to when the georectification is done. Today remote sensing application projects normally involve multi-source datasets, comprising images acquired by different sensors and on different dates. The demand for georectification to ensure all the datasets conform to a standard coordinate system thus often overrides any concern over potential (and subtle) degradation of image information and so georectification is almost always considered the first step in the production of thematic maps from remotely sensed imagery. In this context, it is far more efficient to georectify the raw data so that all derivative images are also georectified.

Though image co-registration and mosaicing can be performed between images based on the GCPs of local matching features, as discussed in Chapter 9, this process becomes redundant once the images are all precisely georectified to the same map projection system. Georectified images of the same area are in fact co-registered while adjacent images of different areas are in a mosaic based on a frame of the map projection coordinates.

### 19.1.2 Image processing

We suggest consideration of the image processing of remotely sensed data in two threads: spectral information and spatial information, as described in Table 19.4. As an example, this procedure may not cover every aspect of image processing nor every application, but it serves a useful guide to the essential image processing techniques and of a workable processing strategy. Alternatively, and depending on the nature of the application, only part of procedure shown in Table 19.4 may be required within a particular project (this is demonstrated in the teaching case studies in Chapter 20). Again, we emphasize that a remote sensing application study should be driven by scientific goals or application objectives, and not by any particular processing procedure.

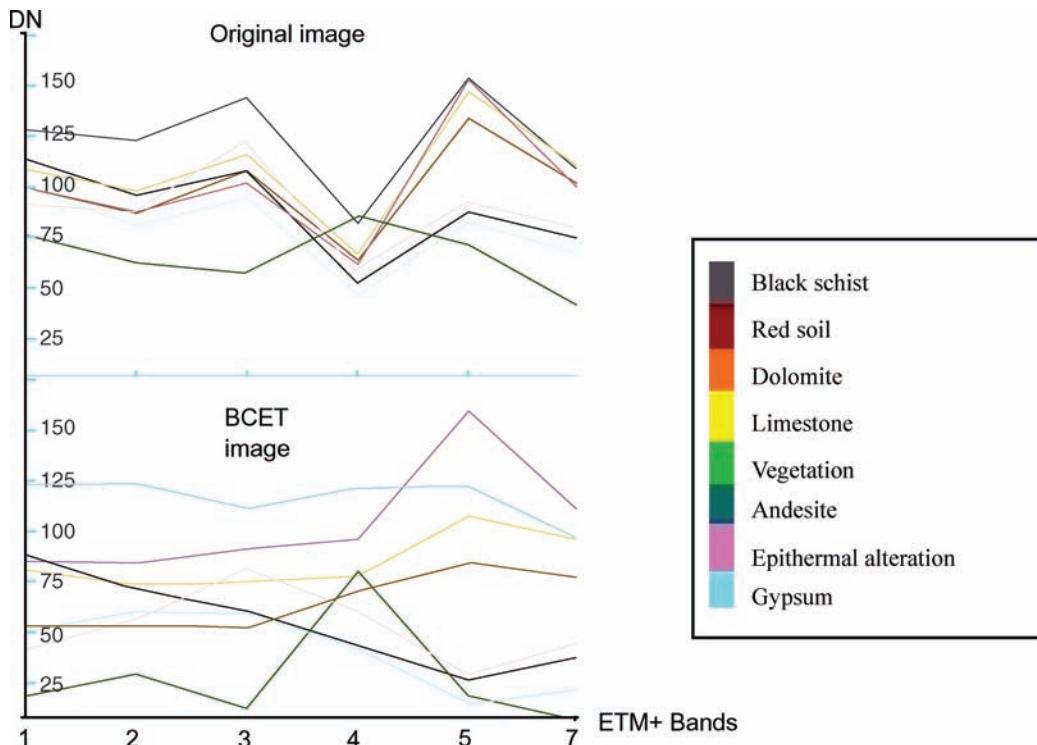
More details of the processing steps given in Table 19.4 are provided in the following sections.

#### 19.1.2.1 Spectral information enhancement and extraction

- *General enhancement for visual observation:*
  - *Optimal contrast enhancement:* Piecewise linear stretch and BCET are preferred but the specific choice and configuration of contrast enhancement techniques should be decided by observation of the image histograms.
  - *False colour composition:* (consider proper band selection based on common spectral signatures). As shown in Figure 2.9 in Chapter 2, for an area with considerable spectral variety, BCET automatically produces an optimal composite image with balanced colours. Piecewise linear stretch enables the generation of a good colour composite interactively.
  - *Decorrelation stretch:* DDS, IHS or PCA decorrelation stretch. As shown in Chapter 5, a decorrelation stretch increases the colour saturation without altering the hues, making ground objects of different spectral signatures more distinctive for visual interpretation.
- *Spectral analysis:* Spectral analysis for target identification is the ultimate goal of hyperspectral image data processing, while for broadband multi-spectral images, the purpose of spectral analysis is to analyse the spectral differences between targets and thus to produce algorithms for selective enhancement and effective target discrimination. In this case, whether or not the

**Table 19.4** Sample procedures illustrating the two component threads of image processing, spectral and spatial, within remote sensing applications

Image Processing for Remote Sensing Applications	
<i>Spectral Information</i>	<i>Spatial Information</i>
General enhancement for visual observation	Data fusion to improve spatial resolution
Selective enhancement	Filtering: Low pass
Enhancement based on data structure and physical models	High pass: Gradient Laplacian
Image classification and segmentation	Spatial component extraction – textural properties, image segmentation and feature extraction
Spectral analysis	



**Figure 19.1** Landsat-7 ETM+ spectral profiles of water, vegetation, red soil, gypsum, mica schist, andesite extracted before (a) and after (b) BCET stretch

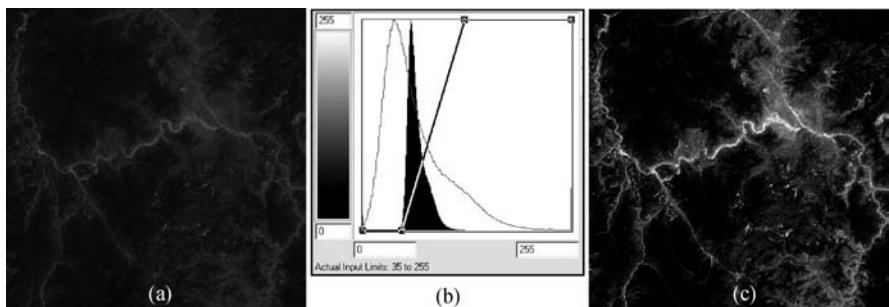
image spectral profile of a target matches its true spectral reflectance signature is not important. What is important is to maximize the differences between different targets. BCET is a simple and effective process for this purpose. As shown in Figure 19.1, the ETM+ spectral profiles of several rock types derived from the original image are very similar in form, with peaks in bands 3 and 5 and a sharp trough in band 4 because they are modulated by very high DN averages in bands 3 (142) and 5 (145) and very low DN average in band 4 (77). BCET balances each band to the same DN average (110) and therefore enhances spectral differences; the spectral profiles of the same rock types derived from the same locations in the BCET dataset of this ETM+ image are distinctively different.

- **Selective enhancement:** Based on spectral analysis (pixel profile or laboratory spectral measurements), selective enhancement algorithms can be composed to highlight or segment specific targets, such as vegetation, water, red soil and

clay minerals. Ratio and difference are the simplest selective techniques. The typical approach to enhance a target is to use the band of the highest DN (reflection peak) to subtract or divide by the band of lowest DN (absorption trough). The differential and ratio indices of vegetation, iron oxide and clay minerals introduced in Chapter 3 (Section 3.5) are all based on this simple principle. One may also consider compound difference and ratio images, using the summation of bands of two peaks against the summation of two troughs, in the same way. Indeed, you may create highly complex algebraic operations, and with good mathematical logic, but do not get lost in doing so!

- *Enhancement based on data structure and physical meaning:*

- **Atmospheric correction:** Atmospheric scattering effects add a constant to multi-spectral images making them look hazy. The spectral bands of shorter wavelength, for example blue and green bands, are more severely affected



**Figure 19.2** (a) This ETM+ band 1 image is rather pale and hazy because of the added constant of atmospheric effects. (b) The atmospheric effects are shown in the solid histogram as the gap between the minimum DN, 35 and 0. Automatic 99% clipping using a piecewise linear stretch effectively removes the added constant of atmospheric effects and stretches the image histogram to fill the display range of 0–255, as shown by the line-delineated histogram. (c) The resultant image shows significantly enhanced contrast with haze completely removed

than those of longer wavelength. Removal of this constant can significantly improve image contrast, and thus quality, and also refine the functionality of ratio technique for topography suppression. There are many techniques for atmospheric correction. The simplest, crude correction is to shift the minimum of an image histogram to zero by clipping or piecewise linear stretch as illustrated in Figure 19.2. The operation is equivalent to the well-known ‘dark pixel subtraction’ technique, proposed by Chavez (1989), but performed more efficiently.

- *PCA and eigenvector analysis:* As fully discussed in Chapter 7, PCA is based on the statistical structure of multi-spectral image data. The analysis of eigenvectors can tell us how each PC image is composed from the spectral bands of the original image and thus how a particular target will be highlighted in a particular PC. This comprises the so-called ‘FPCS’ technique, one of the most effective techniques for PCA-based selective enhancement.
- *Simulated reflectance:* As introduced in Chapter 3 (Section 3.7), we can derive simulated reflectance and thermal emittance from a multi-spectral image with both reflective spectral bands and thermal bands, such as TM/ETM+ and ASTER, based on a simplified model of solar radiation to the Earth. This technique suppresses topography and enhances the spectral signature of ground objects

according to their true spectral reflectance. Thus a simulated reflectance image is directly comparable with its corresponding spectral band and is easy to interpret. In contrast, other techniques, such as ratio and differencing, enhance targets’ spectral signatures indirectly in a combination of two or more bands. While the simulated reflectance technique is for direct enhancement of spectral reflectance of individual image bands, ratio and differencing techniques are for selective enhancement of a particular target on the basis of its spectral properties among several spectral band images.

- *HRGB:* This technique, introduced in Chapter 5 (Section 5.4), is the most effective method of suppressing topographic shadowing and of condensing the spectral information from up to nine spectral bands into a three-band colour composite. The HRGB image does not facilitate easy visual interpretation without reference to simple colour composites because the spectral properties of ground objects are indirectly presented. It is, however, very good for classification, but caution must be taken: since the hue value is in the  $2\pi$  range of a colour wheel, those colour vectors with hues around 0 and  $2\pi$  are numerically very different but are in close proximity in the RGB colour cube.

- *Image classification and segmentation:* Multi-spectral images, as well as multi-source datasets,

may be treated as one multi-variable dataset and so statistical classification algorithms can be applied to produce classification maps, automatically or semi-automatically, from them. Many image features relating to geology and environment can be easily picked up ‘by eye’ on the basis of our knowledge-based understanding of spatial patterns and spectral properties, while many of the tasks that appear to be easy actually turn out to be ‘mission impossible’ for classification. Therefore, image classification cannot replace visual interpretation and, quite often, a classification image may still need visual interpretation. For those ground objects that can be discriminated or identified purely by their spectral properties, classification is the most effective way of mapping them:

- *Thresholding*: This simple technique is very effective for highlighting particular targets. For instance, a thresholded *TMS/TM7* ratio image can reveal those pixels representing hydrated alteration minerals that may indicate the presence of mineral deposits.
- *Statistical classification*: Reiterating what we learned in Chapter 8, supervised classification is based on image training and is often guided by user knowledge that may be biased or incorrect. Unsupervised classification, while sounding like and often appearing like an automatic technique in many image processing software packages, is not one. An unsupervised classification image requires interpretation and can be significantly affected by the choice of initial parameters (most software packages provide only default values!) even though the algorithms have self-optimization functionalities. For both classification approaches, the statistical decision rules dictate the classification accuracy.
- *Spectral angle classification*: The greatest advantage of this algorithm is that classification is independent of any illumination variation or, in other words, it is not affected by topographic shading. Though performed on multi-spectral imagery, i.e. raster data, it is essentially a vector classifier which treats each pixel of an *N*-band image as an *N*-dimensional vector in *N*-dimensional feature space.

### 19.1.2.2 Spatial enhancement

- *Image fusion*: Fusion of a lower spatial resolution colour composite with a higher spatial resolution panchromatic image can generate an apparently higher resolution colour composite that combines both the high spatial resolution of the panchromatic image and the higher spectral resolution of the colour composite. However, we must realize that the image fusion, no matter which technique is used, does not improve the spatial resolution of spectral information in the original colour composite. The processing is for visual observation and interpretation but not for quantitative analysis. In Chapter 6, we introduced the following three fusion techniques:
  - *SFIM*: This is a spectral preservation fusion technique; it maintains the fidelity of the spectral information from the original colour composite. In other words, it does not introduce colour distortion. The technique is, however, very sensitive to image co-registration accuracy. Any co-registration errors may produce subtle edge-blurring effects which degrade the image sharpness.
  - *IHS fusion*: This is achieved by intensity replacement (with a higher spatial resolution image) in an intensity, hue and saturation coordinate system. It is insensitive to misregistration and thus produces sharply enhanced fusion images even if the refined textures actually mismatch with the spectral edges. This technique inevitably introduces distortion of colour and albedo. This distortion will be severe if the spectral range of the three bands forming the colour composite is very different from that of the higher resolution image used for intensity replacement.
  - *Brovey transform*: This is based on direct intensity modulation. It has the same merits and drawbacks as the IHS fusion technique, in terms of fusion quality, but is more efficient in processing as it does not require forward and inverse RGB–IHS transformations.
- *Filtering*: As a neighbourhood processing technique, filtering may not ‘honour’ the original intensity information of an image. Instead, it brings out the spatial relationships between a pixel and its neighbours. On the other hand, all

images acquired through an optical system are, to a degree, filtered images.

- *Low-pass filtering*: The main objective of this is to remove noise at the cost of spatial resolution, but there are many edge-preserving low-pass filters that reduce noise with minimal spatial information loss. For smoothing a classification image, only those filters without numerical calculations should be used.
- *Gradient filter*: As a first-derivative-based filter, this performs directional enhancement. We can configure a gradient filter to enhance the linear features in a particular direction. Important advice in this context is that if linear features in a direction are already very obvious in the unfiltered image, there is no need to apply a gradient filter to enhance this direction at all because it will also enhance subtle features and thus dilute the already clear lineaments.
- *Laplacian filter*: As a second-derivative-based filter, this is non-directional. It enhances image textural edges in all direction. A Laplacian filter is often the first step for texture extraction. One variant of the Laplacian filter is the ‘sharpen’ filter which is equivalent to adding a Laplacian filtered image (textures) to the original image. The result is an edge-sharpened image. Such processing is for visualization and can aid visual interpretation.
- *Spatial component extraction*: Based on neighbourhood processing, many spatial components of a raster dataset may be derived and extracted, such as local contrast, local variance and edge intensity. With DEM data, many of the extracted spatial components have specific physical significance, for instance slope angles, slope aspects, slope curvature and flow grids. These properties relating to surfaces have already been described in Chapter 16 and will be further discussed in the next section.

### 19.1.3 Image interpretation and map composition

When we interpret images to map particular thematic information, either manually or using software annotation tools, we are effectively working on a vector layer. The primary advice here is to start

from the easiest and most obvious and to work towards the most difficult and complex.

Progressing from images to thematic maps is where image processing and GIS merge. As a complete processing cycle, we briefly describe image interpretation and map composition here as the final stage of image processing. We will revisit some contents in this part in greater detail in the next section. Thematic mapping for various application areas may be different but all follow a generic route. The list below describes a general procedure for geological and environmental mapping but it is generically applicable to the mapping of other thematic information:

- *Map format*: Map page setup, geographic coordinates, scale bar, north arrow, title, legend and other relevant general annotation form a standard template for image interpretation and map composition. These are common and standard tools in image processing and GIS software suites, and while they can be applied and modified at any stage of the work, issues such as map scale may greatly affect the detail at which the interpretation can and should be performed, and so should always be considered at the start. Bear in mind that it is considered good practice to capture data at the greatest detail possible, right from start, since it is far easier to reduce detail than to add or recapture it later on.
- *Basic geographic (cultural) information*: Well-enhanced colour composites can provide more than adequate information for the interpretation of:
  - *Human-made features*: e.g., cities, towns, roads, railways and cultivated areas.
  - *Major drainage systems*: e.g., rivers, coast and water bodies.
- *General interpretation*:
  - *Separation of land and sea*: Thresholding using infrared bands often enables the masking of sea pixels from an image, thus enabling more effective enhancement of particular land objects. Be careful, however, to be very critical when choosing the threshold and be aware that dark shadows may cause you problems in this respect.
  - *Vegetation*: Standard false colour composites and vegetation indices enable interpretation of vegetated cover, whether natural or otherwise,

and so are important sources of land use and/or land cover information.

- *Identify major land cover categories:* Agriculture, forest and urban areas (industrial and residential).
- *Geological and environmental interpretation:* This stage forms a dynamic process involving image processing and interpretation. Specific information can be extracted from images produced using purposely designed techniques, as described in Section 19.1.2. From start to finish, the interpretation becomes progressively enriched:
  - Separate the bedrock (solid geology) from any superficial deposits (drift geology) – use both spectral and textual information to do this, such as simple colour composites or edge-sharpened colour composites, or colour composites involving fusion of high spatial resolution. Examine different band combinations and look for targets which may be spectrally similar but texturally different, or vice versa.
  - Interpret major rock types – again, use colour composites of differing band combinations, perhaps with DDS applied to maximize variations. Use PCA to identify areas which are spectrally distinctive, and then try to establish why. Derive simulated reflectance to enable comparison with laboratory data. Produce spectrally complex images, such as via HRGB, to try to distil the spectrally significant information.
  - Highlight specific targets – use differences and ratios, PCA and FPCS to highlight particular target materials. Be careful here to remember that if you have success in highlighting a particular target which you think may be significant, you must also be able to explain why the particular technique has been successful. For instance, discrimination of hydrated minerals and red soils using TM/ETM+ difference images of bands 5 – 7 and 3 – 1 must also be related back to the known spectral signatures of these materials, in order to understand how they work.
  - Interpret structural features – interpretation of lineaments, faults, fractures and fold axes can, in many cases, be achieved without filtering, so filters should be applied only when such features are not clear or obvious. In some

cases, faults and folds are revealed not by their textural characteristics alone, but by the combination and spatial context of spectral variations and topographic shading. In fact, the routine extraction of image ‘lineaments’ should be avoided since these often have no sub-surface geological basis.

- Use classification – this can be quite effective for lithological mapping based on spectral signatures of rocks and minerals, when these are clearly visible at the surface, such as where exposure is continuous. Classification ‘falls down’ where there are irregular spatial patterns, such as those caused by intense faulting/fracturing or where anthropogenic features (cultivated land patterning) are present. One could argue that in such cases where rock exposure is continuous, good ‘old-fashioned’ interpretation of geology is more reliable. It may prove useful when working with very large areas and/or where unknown spectral variations of unknown significance exist, which might not be in the interpreter’s knowledge base and so might otherwise be overlooked. It is our experience that most experienced geologists prefer to avoid classification altogether when interpreting images.

Completion of the composition of a thematic map in this way may be the final stage of a remote sensing application project but more often it forms the beginning of the GIS ‘modelling’ or spatial analysis part of a project, which may have a much wider scope, may involve data from widely different sources and may include other remotely sensed images acquired for very different purposes. In other words, the image-based geological interpretations we have described here may form a tiny part of a much larger and broader project remit.

## 19.2 Remote-sensing-based GIS projects: from images to thematic mapping

Projects involving the creation of thematic maps from remotely sensed imagery commonly follow a similar path, from the simplest background information gathering to the more advanced spatial

modelling. The case studies described here all follow this path and so, from our experience, we have summarised the steps involved into the following general tasks:

- Important preparatory considerations – the regional context and setting of the area being studied. An understanding of the wider context helps to anticipate the variety and types of targets that will require interpretation. This also includes the climate of the area. Tropical and temperate regions will suffer from cloud cover and data acquisition may be problematic. These areas are also likely to be densely vegetated and, while this is not a problem if land use and agriculture are the applications of interest, it will limit the depth and detail of any geological interpretation. Images acquired in tropical areas, even if cloud free, will suffer from haze in the visible bands, which will require correction, and in some cases may render the first three bands unusable for image processing and interpretation. Arid and semi-arid areas make interpretation of ground objects relatively easy since the spectral signatures of the rocks are less likely to be obscured by those of vegetation and thick soils. In some desert environments, wind-blown deposits may also obscure the bedrock and hinder the interpretation of the solid geology beneath.
- A suggested generic procedure – this will always follow the same scheme of three broad phases. The first phase will begin with problem/objective definition, data acquisition/collection, followed by data integration, image processing and analysis, then by interpretation and end with map production and output. This phase should then be followed by fieldwork to verify the results (phase 2). Phase 3 should then involve a refinement of the processing and interpretation, in response to the additional knowledge gained during fieldwork, to arrive at a more complete and realistic interpretation map.
- Mapping using thematic layers derived from remote sensing – data integration and visualization.
- GIS ‘modelling’ based on multi-source data – this demands the integration of data as described above. The point of this exercise is to incorporate different and complementary datasets, in an

attempt to describe or model some potentially complex phenomenon:

- Multi-source data integration and spatial analysis – involving other data, such as geological maps, geophysical data and geochemical data and DEM, which have themselves been processed, interpreted or classified in a particular way which leads to the identification of some complex objective, such as a hazard assessment or a site selection exercise.

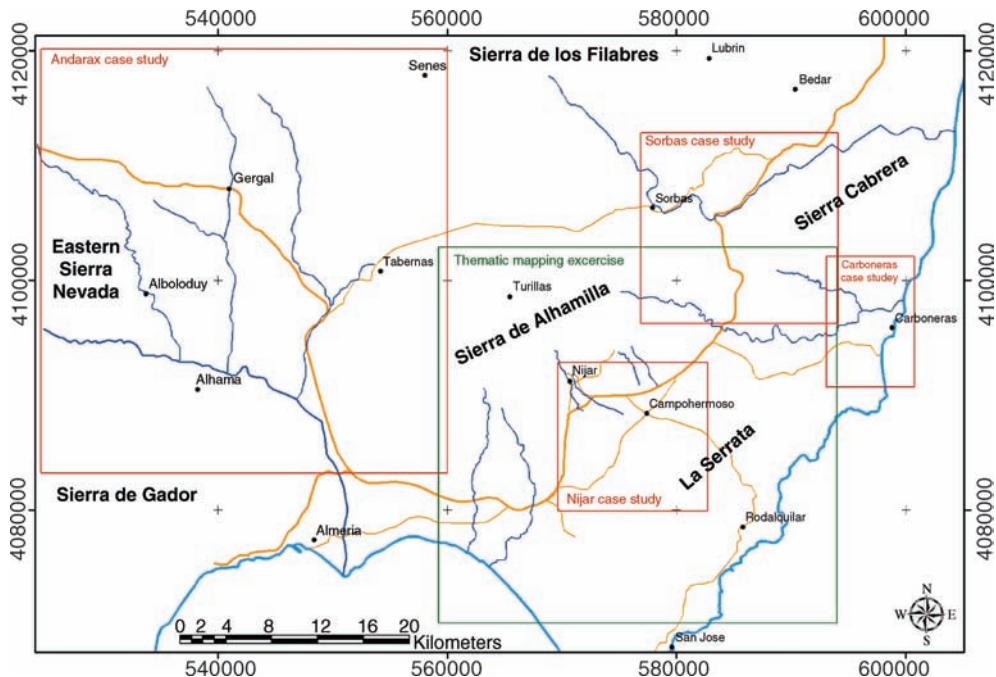
## 19.3 An example of thematic mapping based on optimal visualization and interpretation of multi-spectral satellite imagery

A real remote sensing application project does not necessarily involve all the image processing and GIS modelling as described in the two previous sections. It is always important to remember that for an application project, it should be application driven rather than processing driven. In this section, we present a simple example to demonstrate the use of basic image processing techniques and GIS map composition functionality to produce a digital geological interpretation map. Although the theme is geological mapping, the general approach is applicable for visualization and interpretation of images for other themes too. This case study has been set up as coursework for an introductory course in remote sensing and GIS as part of our undergraduate teaching schedule.

### 19.3.1 Background information

#### 19.3.1.1 Study area

The study area lies in Almeria Province in south-east Spain and is illustrated in Figure 19.3. The environment is characterized by a semi-arid climate (it is Europe’s only semi-desert), sparse natural scrub vegetation, localized intense (covered and irrigated) horticulture, economic extraction of gypsum (and other materials). It is a well-known area for the teaching of field geology, geomorphology and geography. Being semi-arid makes the area ideal for this type of exercise, since there is little



**Figure 19.3** Location map of the Almeria region showing the positions and extents of this thematic mapping exercise (green box, Section 19.3) in addition to the four Spanish case studies described in Section 20.1 Sorbas, Section 20.2 Carboneras, Section 20.3 Nijar and Section 20.4 Andarax (red boxes). Major towns, rivers and coast: fine and bold blue lines respectively; main roads and motorways: fine and bold brown lines respectively

soil development and almost no vegetative cover to obscure the geology.

Despite the limited extent of the area, its geology is varied, which is one of the reasons why it is so popular for teaching. Superficial deposits consist of Quaternary palaeo-alluvial fans and red soils, and Holocene alluvial fans and gravels in ephemeral river channels. The solid geology consists of Palaeozoic graphitic and garnet–mica schists, later Permo-Triassic phyllites and dolomites (in thrust sheets). These form a series of basement massifs between which are sedimentary basins which have been filled by a variety of facies of Messinian sediments, including reef limestones, marls and gypsum. To the south-east, these basement–basin terrains are separated from the Cabo de Gata, a volcanic terrain of the same age as the basin sediments but which has been transported to its current location. This volcanic terrain is typified by acid-intermediate, island-arc volcanics (andesites and dacites), some of which have been subject to late-stage epithermal alteration and mineralization.

The area has quite a long history of gold exploration and mining operations near the town of Rodalquilar (now closed).

Tectonically, the area is still active and structurally it is more complex than at first it may appear. There are major faults, such as the Carboneras fault (a conservative plate boundary), numerous minor accommodation structures synthetic and antithetic to this, in addition to a great many minor neotectonic faults.

#### 19.3.1.2 Data

A Landsat-7 ETM+ sub-scene acquired in June 1999 has been used. The data are georectified to map projection UTM (zone 30N) and data WGS84, and could be said to provide a ‘GIS-ready’ mapping base layer. From this dataset, a series of enhanced colour composites are produced, as information sources for the interpretation of geological features. In addition, a regional geological map, at a scale of 1 : 200 000, and a shaded relief image derived from SRTM 90 m resolution DEM data, are also provided for reference.

Both the latter two datasets provide information at scales which are much coarser than the images being used, and this sometimes causes a little initial confusion on the part of the students. They are only beginning to learn about mapping scales and acceptable/workable levels of detail for (i) discrimination and (ii) feature extraction, and do not at this stage have the conviction (or experience) to know that what they will achieve is a map which should be more detailed, more up to date and positionally more correct than the published regional map. Neither do they realize that the reason for performing this exercise, in reality, may be because geological map information does not exist at the required scales for a particular area, so that mapping from remotely sensed data sources may be the only way. Either way, in a real case, you would always collect as much background information (publications and maps) as you possibly could, to equip yourself, but when it comes to processing the images, you should try very hard not to allow your interpretation to be biased by that information; a great deal of the value of a remotely sensed image interpretation lies in its independence.

### 19.3.2 Image enhancement for visual observation

For this project, the objective of the image processing is quite simple: to produce a few good colour composites from the given Landsat-7 ETM+ image data.

Colour composites of Landsat ETM+ image bands 321, 432 and 531 (RGB) are recommended for general purpose visual interpretation of all ground target types, natural and anthropogenic. The band 321 RGB true colour composite shows the land surface, similar to the way we see it with the naked eye and during fieldwork. The 432 RGB standard false colour composite effectively highlights vegetation in red tones but also major ground objects such as water, soils and gross lithological variations. In semi-arid terrain, where the level of exposure is high and soil and vegetation cover are low, the 531 RGB colour composite is almost always the best image for discriminating lithological variations.

We recommend a simple processing procedure: contrast enhancement using BCET (or linear

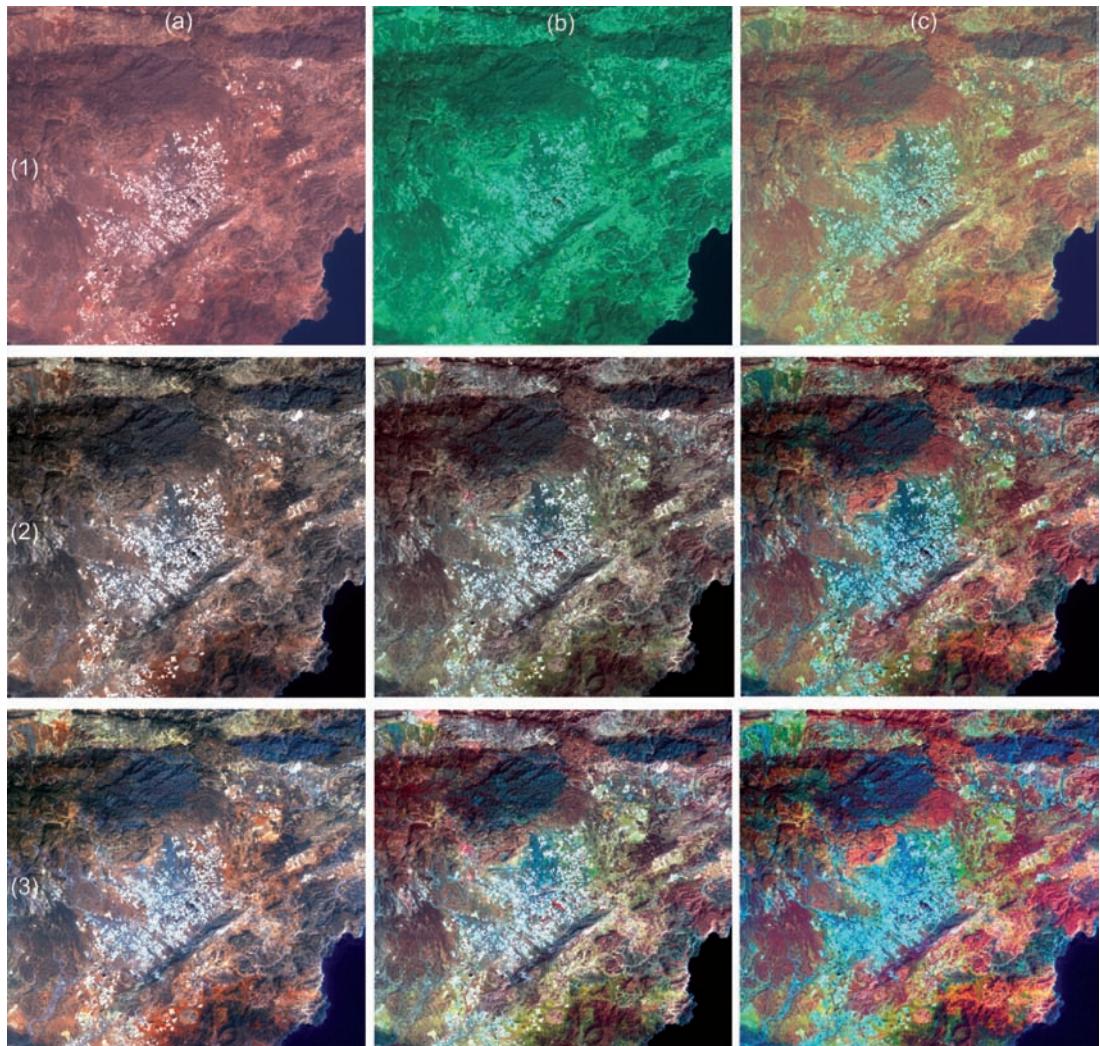
contrast enhancement with appropriate clipping) to balance the colours followed by DDS to increase the colour saturation.

The images shown in Figure 19.4 illustrate the effects of this procedure on the image dataset for this project. The three colour composites are shown in columns (a), (b) and (c), with row (1) representing the raw image composites before contrast enhancement. All three are subject to colour bias to a certain degree and a lack of contrast. After the BCET enhancement, the colour bias in each composite is removed by contrast enhancement and colour enriched, as presented in the second row of Figure 19.4.

Following this, DDS with the achromatic parameter  $k = 0.5$  is applied to enhance further the colour saturation of the three colour composites. These are shown in the third row of Figure 19.4 and present vegetation, red soil, water and drainage patterns, alluvial fans and various rock types vividly in bright and distinctive colours. These three images are then used for visual interpretation in the next stage of the project to generate a digital geological map using GIS.

### 19.3.3 Data capture and image interpretation

You should begin by considering the location of the area being mapped. The most obvious considerations in this case are its regional geological setting and its climate. The former will help you to anticipate the tectonics and lithological characteristics. In this instance, the area lies in the Betic Cordillera of south-east Spain, the rocks range in age from lower Palaeozoic to recent and have undergone two phases of orogenic deformation (Hercynian and Alpine) and the region is still tectonically active. The latter will point to the nature of the terrain surface, whether it is vegetated, weathered to any great depth, subject to persistent cloud and so on. In this case, the area is classified as Europe's only semi-desert (with about 300 mm rainfall per year), there is very little soil development, the atmosphere is hazy from time to time, and there is very little vegetative cover (what vegetation there is, is related closely to ephemeral drainage and irrigation). As a consequence, what is recorded in remotely sensed imagery represents an almost complete record of surface geological exposure across the region,



**Figure 19.4** Landsat-7 ETM+ colour composites of band 321 RGB in column (a), 432 RGB in column (b) and 531 RGB in column (c); and colour composites of the original bands in row (1); after BCET enhancement in row (2); and with DDS after the BCET in row (3)

which makes geological interpretation relatively straightforward. The arid climate makes vegetation a very useful indicator of the presence of groundwater and surface water, and the appearance of localized patches of healthy vegetation usually reveals small rural settlements supported by springs, which are themselves controlled by lithology and structure. Even large-scale agriculture may reveal similar geological control of regional water supplies since this is always more cost effective than piping in water from elsewhere. The aspect of mountain areas will also affect the distribution of

areas (north-facing slopes) that can support natural vegetation and woodland; their presence will need to be considered in interpreting the spectral properties of ground targets. Understanding land use can also therefore be an important factor in interpreting the geology.

Interpretation of different themes in multiple layers:

- Structure of the map project file – the data will likely be organized slightly differently from case to case, because of differences in the specific GIS

software used. Essentially, though, it is sensible to keep solid and drift geological features in separate layers. At this simple level of data capture, it is desirable to capture lithological areas as polygons. This makes for a rather more rapidly constructed map than the more correct method of capturing arcs and later building topology to construct polygons. This choice of strategy rather depends on the time available to complete the task and the software tools available to you. Doing this the ‘quick’, non-topologically correct way means that there are certain limitations on the complexity of information that is captured and conveyed: slivers and gaps, and island polygons will have to be avoided. This method is perfectly acceptable if the final product is required only as a single map product for reference and if no further spatial analysis will be required of the geological polygons.

Other features such as quarries can also be easily stored as simple polygons. Faults on the other hand, by their inherent nature, are stored as linear features in a polyline file. Other *cultural* data can also be captured/imported and stored but should be stored separately from the interpreted features, but could be grouped together for convenience. Such features could include towns (points), roads (polylines) and drainage (polylines). In addition to the images which are the source data for the interpreted features, there are other raster images in the database, namely the SRTM DEM and the regional geological map. Again these raster data layers are, by their nature, stored differently and separately from the vector features, but could usefully be grouped together as reference layers or in two groups, for example satellite images and regional data.

- Use of an interpretation guidance table – during the practical work, students are advised to use a table, such as the one given in Table 19.5, to help familiarize themselves with and note down the appearance of various ground features, as they appear in each colour composite (321, 432 and 531) and to use the suggested legend symbols (or ones of their own making) to annotate their interpretation. This forms an important step in understanding the way in which the displayed spectral bands determine the colour of features in each image. The connection is made between

relative reflectivity in particular wavebands (Landsat bands 1, 2, 3, 4 and 5 in this instance) and image brightness in particular colour bands (red, green and blue). For instance, iron-oxide-rich red soils appear bright red in the 321 (RGB) image but are greenish in both the 432 (RGB) and 531 (RGB) images since the relative reflectance of iron oxides is high in Landsat band 3 and low in band 2 and even lower in band 1.

The items listed in the table also provide as a hint towards the lithologies that the students should expect to see in this area. In this example, the students have visited many localities in the area during the previous year and so are familiar with its geology and geomorphology; they simply need reminding of what they saw and learned. They also need encouragement to make the link between the appearance of rocks in the field and a hand specimen, with their appearance in the image, and to treat the work rather like a complicated puzzle to which there may be no definitive answer.

- Procedure – we always recommend spending some time just looking at the images, familiarizing yourself with the database and software and examining how different targets appear in each image band combination. During this process, you will probably begin conceptually subdividing the study area into geologically significant zones (or terrains), before actually capturing any new data. This will help you to understand which are the most important features to convey in the final map. During this stage you should also establish the optimum scale at which to capture features, according to the spatial resolution of the images, probably around 1 : 35 000 in this case, where you can see maximum surface detail but not individual pixels.

You should soon begin to feel confident about identifying boundaries between objects which are spectrally and texturally different – these will probably be the most obvious and largest, eye-catching features. When you are ready to capture some features, start by capturing those most obvious lithological outcrops, i.e. their boundaries, and when you have created a feature, remember also to enter an identifying, descriptive name for the feature, in its attribute table. Doing this for each feature as you create it will save

**Table 19.5** A sample image interpretation guidance table – an aide-memoire

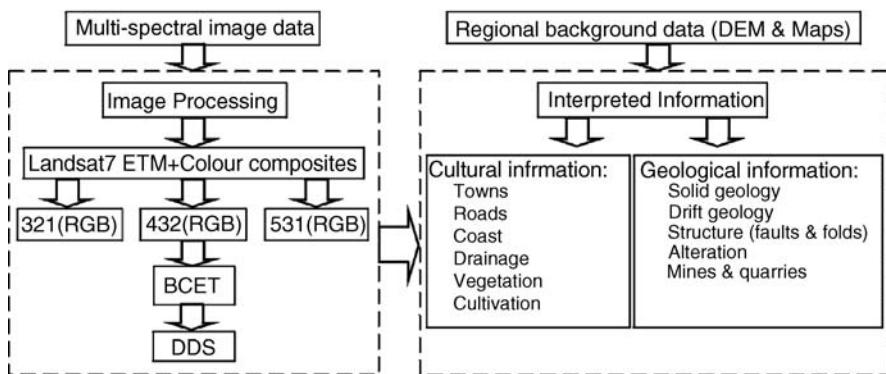
Ground Objects	Description of the features in Landsat images			Legend symbols
	321RGB	432RGB	531RGB	
Natural Vegetation				
Horticulture				
Urban areas				
Quarries				
Alluvial fans				
River debris				
Schists				
Dolomites				
Limestones				
Marls				
Gypsum				
Andesite				
Dacite				
Solid & drift geological boundaries				
Inferred geological boundaries				
Faults (Major)				
Faults (Minor)				

time later (when you may not remember quite so clearly what you were thinking at the time). At this stage you will almost certainly not be able to give the outcrop a specific geological identity, you may have little idea about the lithology, but that does not matter at this stage. You will build up a series of units identified perhaps as ‘sedimentary\_1, \_2, \_3’ or ‘volcanic\_1, \_2’ and so on. As you proceed in this way, moving around the image, you will find outcrops which look spectrally and perhaps texturally similar to some

that you have already captured, so you should soon find that you have several polygons with the same identifying code or label. You can of course amend these descriptions as you proceed.

After identifying all the obvious features, you will then begin to find boundaries between lithologies which are only subtly different, and perhaps spectrally complex in themselves. For instance, they may be spectrally similar but texturally different, in which case you may surmise that they may represent chemically similar lithologies

## Remotely Sensed Image Based Geological Mapping



**Figure 19.5** Work-flow chart, summarizing the processing and interpretation procedures of the mapping project described in this section

which are not the same in terms of stratigraphy or facies. Alternatively, they may be spectrally different but texturally contiguous, in which case you could conclude that they represent local variations of lithology which have common structure, such as suites of metamorphic rocks in mountain ranges which have been subject to regional deformation. As you progress around the image, capturing lithological information, many questions will arise and as you attempt to answer those questions, calling on your own geological knowledge and experience, you will get closer to giving more precise geological names to the outcrops. After some time, it is advisable to stop and do something else, returning later with a fresh and critical eye, to go over what you have done and refine it. You may repeat this process a number of times. During these times, it is always a good idea to zoom back out from your detailed observations, to a more regional scale, since it is very easy to get ‘bogged down’ in the detail and to spend more time on one small area than (i) you can afford and (ii) is necessary for the scale of the map you are creating.

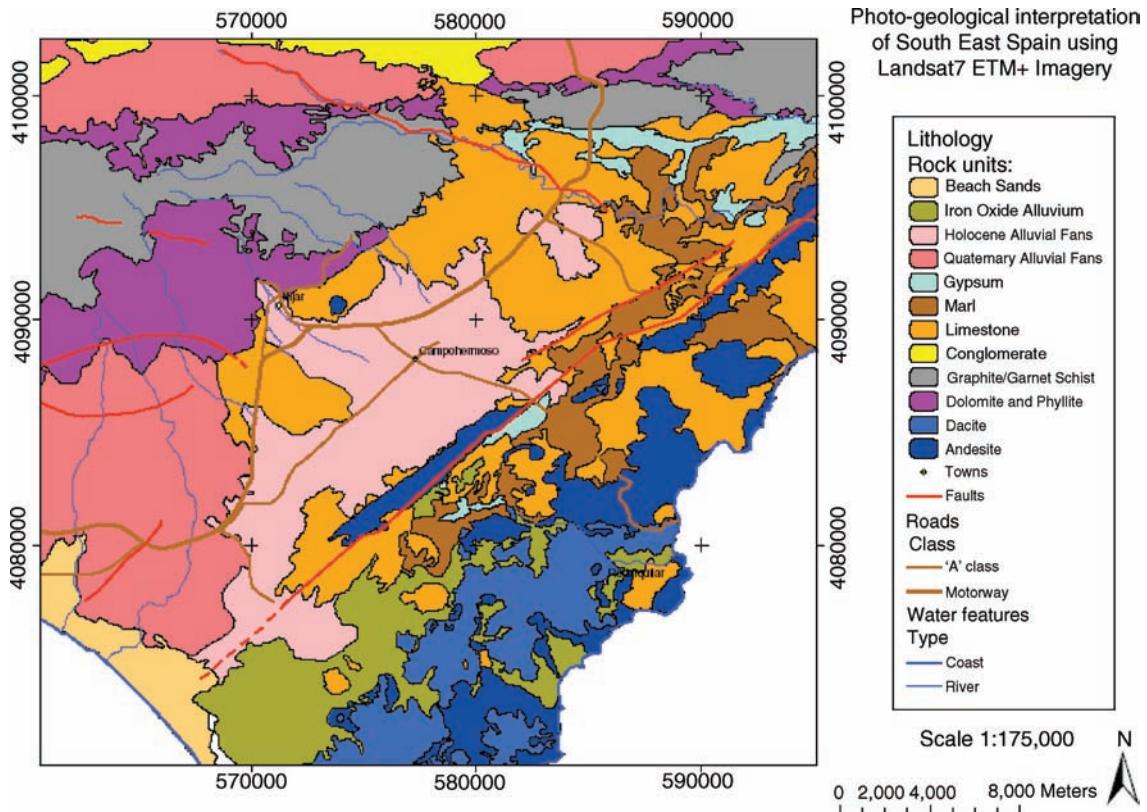
- Structure – a considerable amount of common sense and logic is required to be successful and this comes with experience, confidence and clear thought. You will also, no doubt, find objects and features which you cannot identify or understand at all and these are the ones that should be

recorded in your notebook as features which require verification during fieldwork (Figure 19.5).

### 19.3.4 Map composition

The finished product will include the interpreted geological information plus sufficient cultural information to make the map navigable, and items normally found on any map, such as a coordinate grid, scale bar, north arrow, annotation, title and map legend to explain colours and symbols used on the map.

Care needs to be taken in the selection of colours, particularly for the interpreted polygons. Bearing in mind that in the world of digital mapping using highly functional software a dazzling array of colours and symbols is available, it is important to take a step back and reconsider the objectives considered at the start of the exercise (in Section 19.1). Remember what the application theme and overall objectives are. What aspect of the map should be the most obvious one to the eye of the intended end-user of the map? Consider the map scale for the final map; remember to ensure this is a sensible number (preferably a whole number rather than some obscure fraction) and that all the information you want to show appears sensibly laid out on the page. Make sure that the map is not overcrowded with either cultural information or labels. Will the symbols/labels be discernible/legible in the final scaled version? It is often worth making one or more test



**Figure 19.6** An example of the finished interpretation map

prints of the map to establish this. Remember to consider who might be using the map: consider if it conveys sufficient background/auxiliary information to explain sufficiently what the map shows, how it was produced, what it represents, and so on, to the untrained eye; assume that a layperson will need to make use of it and then assess whether they will actually be able to understand it.

Given that the database contains height data, in the form of the SRTM DEM, both the images and the finished map can then be visualized in pseudo three dimensions.

Theoretically, the finished map should be of potentially publishable standard. In this class exercise, the results will always fall short of that standard but will be an impressive achievement for each student nonetheless, and will look something like the example shown in Figure 19.6.

When the final map is complete, it is then output to one of a number of standard formats, such as

.pdf, .tiff or .jpg, or to some other format which supports zooming and some query functions (such as ArcPublisher), ready for sending to its final destination, wherever that may be. Extremely useful, at this stage, is the ability to output the coordinate reference information with the publishable map, using a *world file*, either a .tfw or .jfw (as described in Chapter 12), so that it can then be displayed in any other GIS as a map database product.

## 19.4 Summary

This is a very important chapter since it sets out a kind of recipe for the most logical way to approach a typical project using remotely sensed data, to achieve an application objective. In this case, we use basic geological mapping as an example but the topic could easily be land use, environment, agriculture or water resources. The important thing

common to all such projects is the strategy. We have thus tried to provide some valuable and important rules of thumb, which we know from our experience of doing this type of work. There can surely be no better way to learn anything than by simply doing it; what we have done here is to accelerate the learning process by steering you away from the many, known, potential pitfalls that lie along the way. From defining the project goals, through to extracting the elusive thing at the very end, namely the real image information, this chapter presents a simple and generic formula for doing so.

## Questions

- 19.1 What other factors govern the choice of image data used?
- 19.2 What are the potential uses of the extracted thematic/vector data?
- 19.3 How much time is required to achieve these steps?
- 19.4 What is/are the most appropriate software tool(s)?
- 19.5 How does this phase of work link with the wider scope of project development?

# 20

## Thematic Teaching Case Studies in SE Spain

In this chapter, we discuss several teaching case studies on specific themes, using image data of SE Spain, to demonstrate remote sensing applications in earth and environmental sciences. Each case emphasizes different parts of the general strategy (described in the previous chapter) but all follow the same route from image processing to information extraction and finally to thematic mapping.

### 20.1 Thematic information extraction (1): gypsum natural outcrop mapping and quarry change assessment

The Sorbas area, in Almeria Province of south-eastern Spain, contains one of the largest and most significant gypsum deposits in the world. The large-scale economic extraction and environmental conservation of the natural gypsum karst landscape are in direct conflict. In this case study, multi-temporal TM/ETM+ images are used to map the distribution of natural gypsum outcrops and to chart the temporal changes in the extents and location of gypsum quarrying, to provide objective information relating to the impact of the extraction industry on

the regional environment. The main objectives of the study are:

- Identify and map the natural outcrops of gypsum.
- Extract gypsum quarries and investigate the changes of gypsum quarries over 16 years.

The study comprises three parts:

1. Multi-spectral image enhancement for gypsum mapping.
2. Gypsum quarry extraction.
3. Multi-temporal comparison for quantitative assessment of the change of gypsum quarries.

#### 20.1.1 Data Preparation and general visualization

Three TM/ETM+ images with eight-year intervals, acquired in 1984, 1992 and 2000 (Table 20.1), are used in this study. The Landsat-7 ETM+ image acquired in 2000 was downloaded from the Global Land Cover Facility (<http://www.glcfc.umiacs.umd.edu/index.shtml>), which has been accurately rectified to UTM N30 based on WGS84 data. The other two images were co-registered to the ETM+ image to conform to the same map projection.

**Table 20.1** The TM/ETM+ images used in this case study

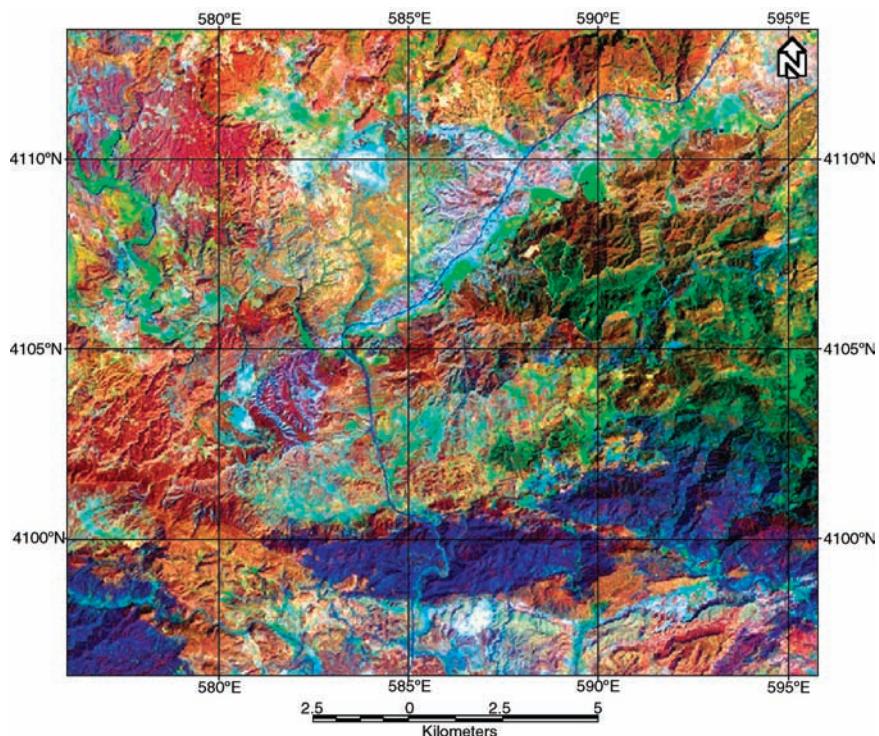
Satellites	Sensors	Path/row	Acquisition date
Landsat-4	TM	199/034	19 July 1984
Landsat-5	TM	199/034	25 July 1992
Landsat-7	ETM+	199/034	8 August 2000

As the three images were acquired by the same type of sensor system and from similar orbits, although onboard different satellites, the major deformations between them are produced by linear translation and rotation. The simplest linear polynomial transform therefore produces the best co-registration accuracy.

As explained in Chapter 19 (Section 19.1.2), to optimize the spectral analysis and visualization,

BCET has been applied to produce BCET datasets corresponding to each of the three images.

Remote sensing can reduce the workload of field investigation significantly but cannot replace it. Field knowledge, of the gypsum outcrops, and existing geological maps are of great assistance in sampling to produce image spectral profiles and to assess the results of this study. Where field knowledge and existing maps are unavailable, an understanding of the spectral properties of major ground objects and target minerals is essential, while general image visualization is the starting point for gaining this knowledge. Figure 20.1 is a colour composite of the ETM+ bands 541 in RGB with BCET and DDS enhancement. The image displays vegetation in green and various rock types in a variety of different colours. The very bright patches in this image are produced by the gypsum quarries. Spectral profiles of gypsum can be sampled in these quarries and nearby areas of outcrop.



**Figure 20.1** Landsat-7 ETM+ bands 541 RGB colour composite with BCET and DDS enhancement to show vegetation, lithology and quarries of the study area. The scale bar in this image serves as a reference to all the images, which cover exactly the same area, in this case study

### 20.1.2 Gypsum enhancement and extraction based on spectral analysis

As a hydrated mineral, gypsum is spectrally similar to clay minerals relating to alteration characterized by high reflectance in TM/ETM+ band 5, a broad SWIR spectral band centred at  $1.65\text{ }\mu\text{m}$ , and strong absorption in TM/ETM+ band 7, a broad SWIR spectral band centred at  $2.2\text{ }\mu\text{m}$ . A simple differencing or ratio of TM/ETM+ bands 5 and 7 can therefore selectively enhance both targets but cannot achieve the separation between them (Figure 20.2). Figure 20.3b shows the ETM+ image spectral profiles of a known gypsum quarry and natural outcrop, an epithermal alteration zone and vegetation. The unique spectral feature of gypsum that makes it different from clay minerals is that it has slightly higher reflectance in band 4 (nearer infrared) than in band 5, where alteration clay minerals have a strong absorption trough. Thus, a simple BCET DDS colour composite of TM/ETM+ bands 4, 5 and 7 in RGB highlights gypsum uniquely in yellow, separating it from the alteration clay minerals in cyan as shown in Figure 20.3a. The spectral sample points are denoted in this image as well.

The spectral signature of gypsum bears some similarity to that of vegetation in TM/ETM+ bands 4, 5 and 7 as shown in Figure 20.3, but the so-called ‘red-edge’ feature of strong absorption in the red band (TM/ETM+ band 3) in contrast to very high

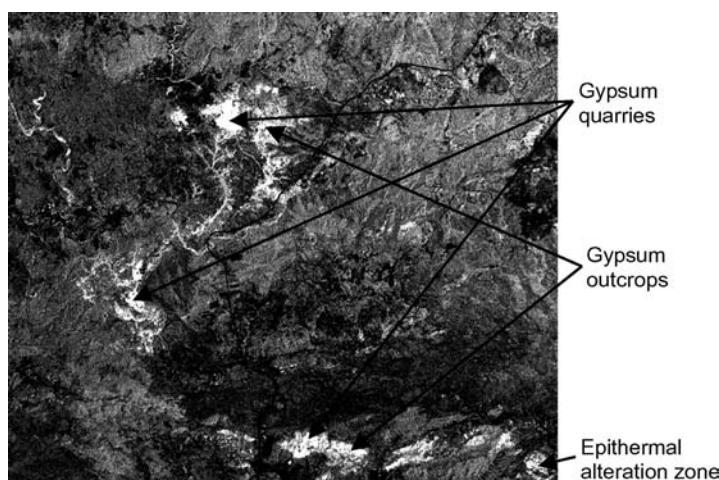
reflectance in the NIR band (TM/ETM+ band 4) is unique to vegetation. With these observations of spectral signatures in mind, a simple technique is designed to extract gypsum with the following algebraic and logical operations:

$$\begin{aligned} \text{If } \frac{TM4 - TM3}{TM4 + TM3} > 0.1 \text{ then } 0 \text{ else} \\ \text{if } TM4 - TM5 \geq 0 \text{ then } TM5 - TM7 \text{ else } 0. \end{aligned} \quad (20.1)$$

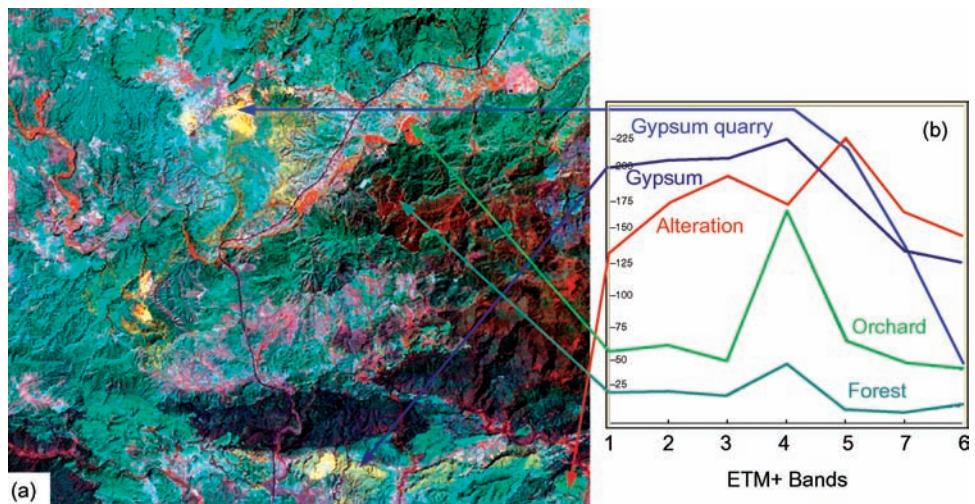
In this formula, the first condition is the NDVI to eliminate vegetation and the second condition  $TM4 - TM5 \geq 0$  is to exclude clay minerals. Thus gypsum (both outcrops and quarries) is extracted in a single image with a threshold  $TM5 - TM7 \geq 10$  as shown in Figure 20.4.

The key difference between the gypsum quarries and natural outcrops of gypsum is the very high albedo of the smooth quarry floor in visible spectral range (Figure 20.3b). Thus a slight modification of formula (20.1) to add the red band in the final operation will extract the gypsum quarries only:

$$\begin{aligned} \text{If } \frac{TM4 - TM3}{TM4 + TM3} > 0.1 \text{ then } 0 \text{ else} \\ \text{if } TM4 - TM5 \geq 0 \\ \text{then } TM5 - TM7 + TM3 \text{ else } 0. \end{aligned} \quad (20.2)$$



**Figure 20.2** Difference image of ETM+ band 5 minus band 7. Most white patches are either gypsum quarries or gypsum outcrops except the one in the bottom-right corner, which is an epithermal alteration zone



**Figure 20.3** (a) Colour composite of ETM+ bands 457 RGB with BCET and DDS enhancement. Gypsum outcrops and gypsum quarries are uniquely highlighted in yellow colours while vegetation is in red and reddish orange, and alteration clay minerals are the same as several rock types in cyan; and (b) ETM+ image spectral signatures of gypsum quarry, outcrop of gypsum, alteration zone, orange orchard and pine tree forest. The arrows indicate the spectral sample position in the image

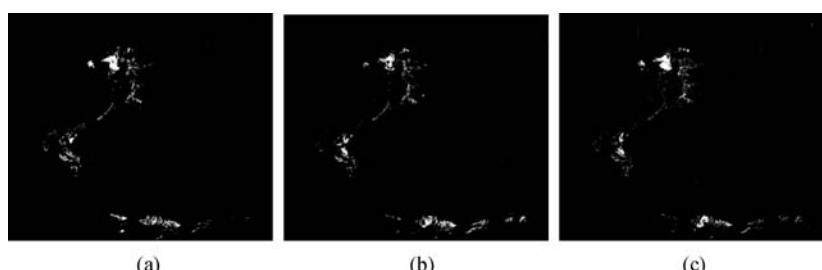
The extracted gypsum quarries in 1984, 1992 and 2000 are presented in Figure 20.5. As shown in each of the corresponding histograms, a threshold is set in the bottom of the trough between the main peak on the left and a small hump on the right that represent the quarry pixels. The threshold sets the DNs of quarry pixels to 255 and others to 0.

### 20.1.3 Gypsum quarry changes during 1984–2000

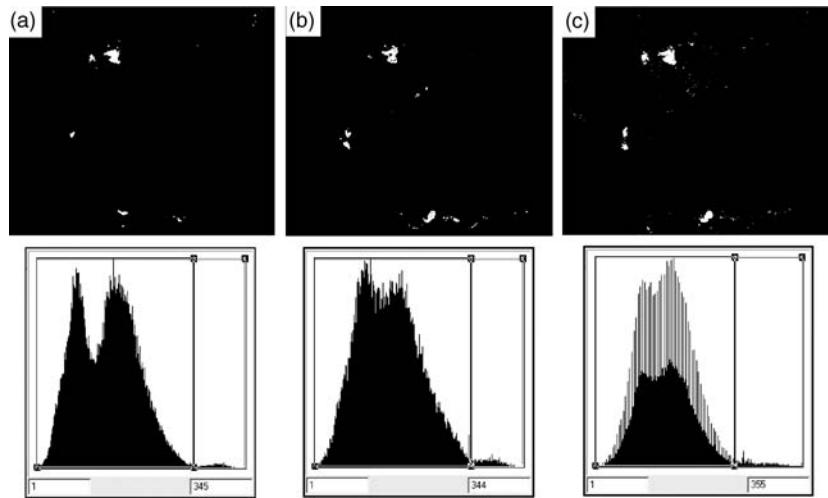
We can display the gypsum quarry extraction images of the years 2000, 1992 and 1984 in red, green and blue to formulate an RGB colour composite

as shown in Figure 20.6. The colours of the extracted quarries in this colour composite indicate the temporal change and development of gypsum quarrying in the region as interpreted in Table 20.2.

Quarry 1 is the largest in the image. The patch is mostly white, indicating that the quarry was already on a great scale in 1984 and in operation throughout the following years to 2000. The surrounding yellow belt along the east and north margin is the quarry expansion during 1984–1992, while the red belt surrounding the south half of the quarry indicates the quarry expansion during 1992–2000. The green patches on the north edge of the quarry are interesting, and denote the areas quarried during 1984–1992 and then the ground



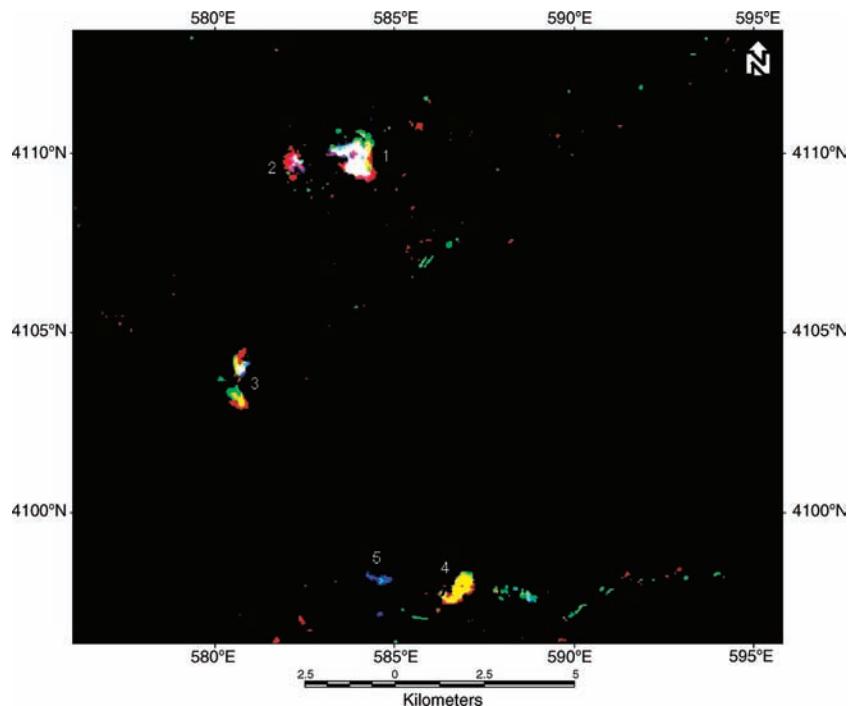
**Figure 20.4** Images of gypsum extraction and mapping: (a) 1984; (b) 1992; and (c) 2000



**Figure 20.5** Gypsum quarry extraction images and their corresponding histograms and thresholds: (a) 1984; (b) 1992; and (c) 2000

was probably restored. The small pink patch in the centre of the white area marks the lowest bottom of the quarry which can easily become a water pond after heavy rain as shown in the field photo in

Figure 20.7. The pink colour of this little patch implies that it was filled with water on the date of the 1992 image and was dry on the dates of the 1984 and 2000 images.



**Figure 20.6** Colour composite of gypsum quarry extraction images of year 2000 in red, 1992 in green and 1984 in blue. The colour interpretation is detailed in Table 20.2

**Table 20.2** Interpretation of colours of Figure 20.6

Colour	Interpretation
White	Gypsum quarries before 1984 (in 1984, 1992 and 2000 images)
Yellow	Gypsum quarries after 1984 (in 1992 and 2000 images)
Red	Gypsum quarries after 1992 (only in 2000 image)
Green	Gypsum quarries after 1984 and before 2000 (only in 1992 image)
Blue	Gypsum quarries before 1992 (only in 1984 image)
Cyan	Gypsum quarries before 2000 (in 1984 and 1992 images)
Magenta	Gypsum quarries before 1984 and after 1992 (in 1984 and 2000 images)

Quarry 2 was started before 1984 as indicated by a small white patch in the east of the quarry. Significant development occurred after 1992 to the west as illustrated in red. Quarry 3 was started in the north part before 1984 as well and then saw considerable expansion during 1984–1992, shown in yellow and green, and 1992–2000, in red.

Quarry 4 is in a major belt of gypsum natural outcrops as highlighted in Figure 20.4. Shown in yellow colour, this second largest quarry in the image was started after 1984 and quickly reached the scope recorded in the 1992 image.

The expansion after 1992 till 2000 was limited. There are some blue, green and cyan patches nearby, along the gypsum outcrop belt. These patches are abandoned quarry trials in the years before 2000. In particular, quarry 5 in blue was sizeable before 1984 but the quarrying operation was ended before 1992.

Besides the major quarries described above, there are some scattered isolated dots in Figure 20.6, which are not likely gypsum quarries. These could be some casual diggings of gypsum as well as incorrectly extracted pixels.



**Figure 20.7** Field photo of the gypsum quarry 1 in Figure 20.6 taken in 2003. The lowest part of the quarry has become a water pond with vegetation

#### 20.1.4 Summary of the case study

In this case study, we demonstrated how to design simple and effective image processing techniques to map gypsum outcrops and extract gypsum quarries based on image spectral profile analysis. Though the image spectral profiles of six TM/ETM+ reflective multi-spectral bands and a thermal band are fairly crude in comparison with laboratory spectral profiles, the diagnostic spectral property of gypsum can be enhanced to achieve effective discrimination for accurate extraction of gypsum outcrops and quarries. Usually, broadband multi-spectral image data are only adequate for ground object discrimination but not identification; however, the identification can be easily achieved through minimal field investigation guided by these images of target-oriented enhancement and thematic extraction. A field investigation indeed forms an essential part of this type of case study. All the quarries and natural outcrops extracted in the resultant images of this project had been verified in our field trips for MSc students of remote sensing.

## 20.2 Thematic information extraction (2): spectral enhancement and mineral mapping of epithermal gold alteration, and iron ore deposits in ferroan dolomite

The Carboneras area lies to the south-east of the town of Sorbas, on the eastern coast of Almeria Province, Spain (refer back to Figure 19.3). A regional NE–SW-oriented tectonic system, known as the Carboneras fault zone, cuts the area into complicated jumbled slices of Palaeozoic and Mesozoic basement schists, phyllites and dolomites, together with pockets of Tertiary volcanic rocks.

Intense epithermal alteration has resulted in the enrichment of economic minerals within a small gold deposit in the study area. Some exploration has been carried out but was later abandoned because of the low grade and limited size of the deposit. However, the extensive alteration zone serves as a good test area to demonstrate the application of

multi-spectral remote sensing for mineral exploration. Another mineral of economic interest in this area, which has been actively mined elsewhere, is iron ore, found here within Triassic dolomite. Although closely associated geographically, the ferroan dolomite deposits were accumulated through a quite different process from the iron oxides associated with the epithermal gold deposits; the highly fractured dolomites have become enriched via weathering and leaching (i.e. they are gossan-type deposits).

With two distinctively different mineralization systems within close proximity, we use this case study to demonstrate the effectiveness of simple multi-spectral enhancement techniques for mineral exploration with two objectives:

- Locate argillic–siliceous alteration zone, the Carboneras gold prospect.
- Locate Triassic ferroan dolomite and iron minerals (limonite).

Using the 11 band airborne thematic mapper (ATM) and Terra-1 ASTER 14 band images, the study comprises three steps:

1. Image dataset preparation.
2. ASTER image processing and analysis for regional prospecting.
3. ATM image processing and analysis for target extraction.

### 20.2.1 Image datasets and data preparation

Two images, namely an ATM image taken in 1991 (NERC UK Airborne Remote Sensing Facility) and a Terra-1 ASTER image taken in 2002, are used in this study. The details of the ATM and ASTER sensors in comparison with Landsat TM/ETM+ are listed in Table 20.3. It is important to notice that the three ASTER spectral groups of VNIR, SWIR and TIR are not only at different spatial resolutions but also in different radiometric quantization ranges. The VNIR and SWIR bands are in an 8 bit value range while the thermal bands are in a 16 bit one.

It is typical of airborne image data that the ATM image is subject to various localized geometric distortion caused by the unstable imaging status of the aircraft. As the first step for data preparation, the

**Table 20.3** Comparison of spectral bands and spatial resolution of the ASTER and ATM images used in this case study with the Landsat TM/ETM+

Sensor systems	Terra-1 ASTER			Landsat-3–7 TM/ETM+			ATM (7.5 m)		
	Spectral region	Band	Spectral range ( $\mu\text{m}$ )	Spatial resolution (m)	Band	Spectral range ( $\mu\text{m}$ )	Spatial resolution (m)	Band	Spectral range ( $\mu\text{m}$ )
VNIR							30	1	0.42–0.45
				15	1	0.45–0.53		2	0.45–0.52
	1	0.52–0.60			2	0.52–0.60		3	0.52–0.60
	2	0.63–0.69			3	0.63–0.69		4	0.605–0.625
								5	0.63–0.69
								6	0.695–0.75
	3N	0.78–0.86			4	0.76–0.90		7	0.76–0.90
	3B	0.78–0.86			Pan	0.52–0.90	15	8	0.91–1.05
SWIR	4	1.60–1.70		30	5	1.55–1.75	30	9	1.55–1.75
	5	2.145–2.185			7	2.08–2.35		10	2.08–2.35
	6	2.185–2.225							
	7	2.235–2.285							
	8	2.295–2.365							
	9	2.360–2.430							
TIR	10	8.125–8.475		90	6	10.4–12.5	TM 120	11	8.5–13
	11	8.475–8.825							
	12	8.925–9.275							
	13	10.25–10.95					ETM+ 60		
	14	10.95–11.65							

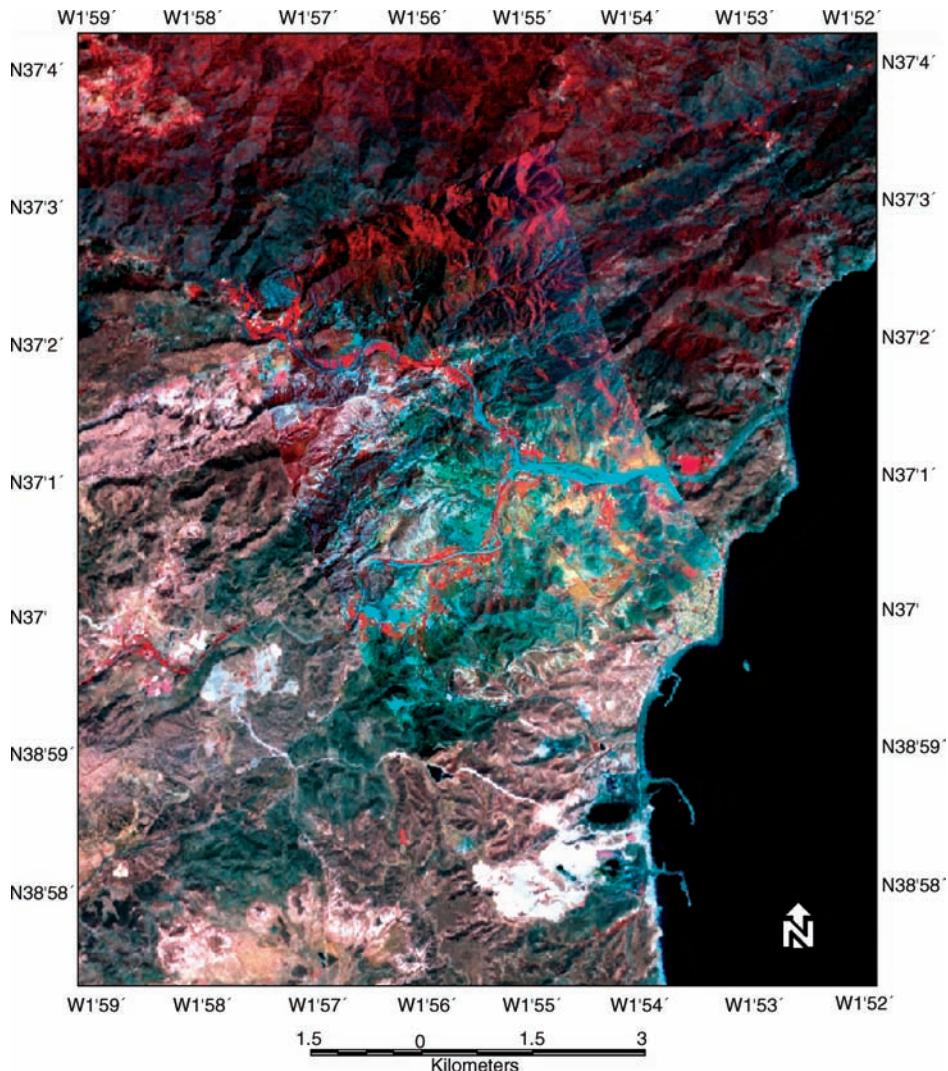
ATM image was rectified to the ASTER image that complies with UTM N30 based on ED50 datum. The warping transformation was a cubic polynomial fitting derived from 25 GCPs and the bilinear resampling was applied to produce the rectified output image from the input image DNs. Because of the significant irregular distortion of the ATM image, the RSM of GCPs ranges from 1 to 22 pixels even though these GCPs were quite carefully selected. We therefore do not expect an accurate co-registration between the two images, which can only be processed and analysed separately for comparison. The rectified ATM image is in a curved irregular shape indicating nonlinear distortion of the image in the reference map projection.

Figure 20.8 is a merged display of standard false colour composites of ASTER (bands 321 in RGB) and ATM (bands 753 in RGB) images. It shows that the rectified ATM image in the middle matches the ASTER image fairly well; however, a closer look

reveals considerable discrepancies between the two. Efforts for integrated processing and image-fusion-based analysis of the two images will introduce more errors than benefits. In this case study, we use the low-spatial-resolution ASTER image for regional prospectivity and the high-spatial-resolution ATM image to focus on the area of interest for target mineral extraction.

## 20.2.2 ASTER image processing and analysis for regional prospectivity

This case study was chosen because we already know the area quite well through image study and previous field investigations. However, if we presume little knowledge about the area but recognize that it might have potential mineralization based on the regional geological setting, then the first step in studying the area would be to conduct regional



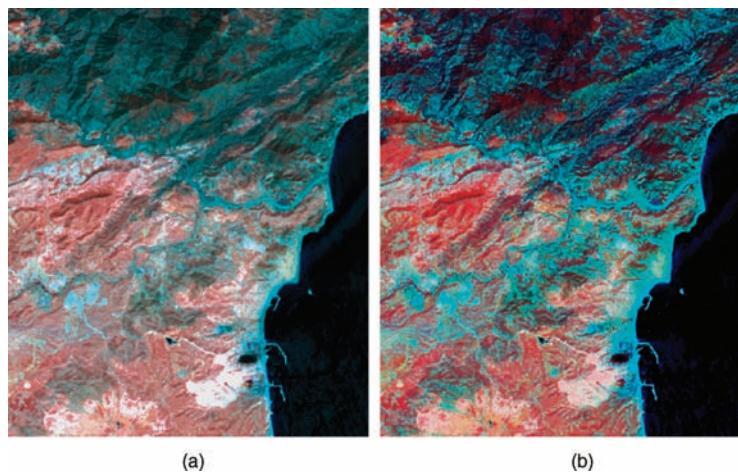
**Figure 20.8** Merged display of standard false colour composites of ASTER (bands 321 in RGB) and ATM (bands 753 in RGB) images

prospecting using a satellite image with adequate spatial resolution and large coverage. The ASTER image with three 15 m spatial resolution VNIR bands, six 30 m spatial resolution SWIR bands and five 90 m spatial resolution thermal (TIR) bands serves the purpose well. The particular advantage of the ASTER image is its very high spectral resolution in the SWIR and TIR bands.

Without knowing the mineral targets, well-established techniques should be tried first, from visualization to general spectral enhancement. Firstly, colour composites of ASTER bands 421 in

RGB with BCET and DDS are generated as shown in Figure 20.9. These colour composites present rich information on general lithology and geological structure (refer to the geological map in Figure 19.6), as well as rivers, quarries and human-made structures, but they do not show obvious features indicating minerals.

As we studied before, simple standard differencing and ratio techniques can effectively locate clay and hydrated minerals. ASTER imagery has six SWIR bands with band 4 equivalent to TM band 5, and bands 5–9 are high-spectral-resolution

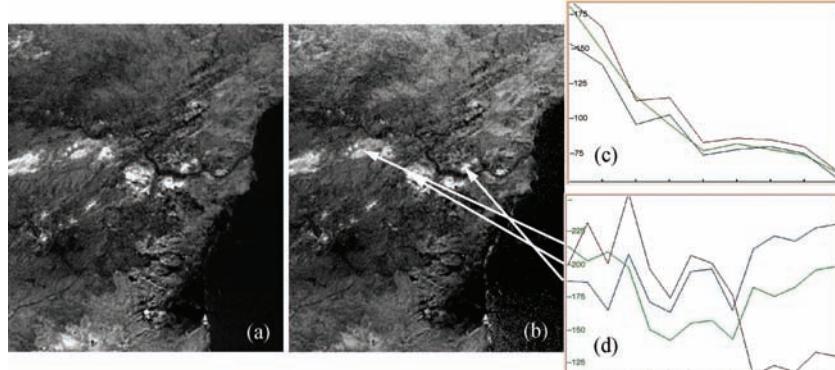


**Figure 20.9** ASTER bands 421 RGB colour composites: (a) BCET; and (b) DDS

bands within a narrow range of  $2.145 - 2.43 \mu\text{m}$  that largely overlap the spectral range of TM band 7. Most alteration-related clay minerals and hydrated minerals have a deep absorption trough in the narrow spectral range depicted by ASTER band 6. We thus expect that the ratio or differencing images between ASTER bands 4 and 6 can highlight potential targets of such minerals. As shown in Figure 20.10, both band 4 – band 6 and band 4/band 6 highlight an obvious bright belt in the middle of the image. The two techniques do not show much difference in the results. Spectral profiles of selected bright pixels in several patches from the original image data appear to have similar shapes to the diagnostic absorption features in VNIR and SWIR

bands 1–9, implying the same type of alteration or hydrated minerals (Figure 20.10c). However, the spectral profiles of the same points from the BCET data of bands 1–14 (the thermal bands were rescaled to the 8 bit value range) in Figure 20.10d are clearly in three distinctive groups:

- Argillic alteration: High reflectance in band 4, strong absorption in both band 6 and band 3, and low emission in thermal bands 10–14.
- Siliceous alteration: Similar spectral signature to the argillic alteration in VNIR and SWIR bands but the high emission in thermal bands 10–14 is diagnostic for underpinning quartz and silica-rich minerals.



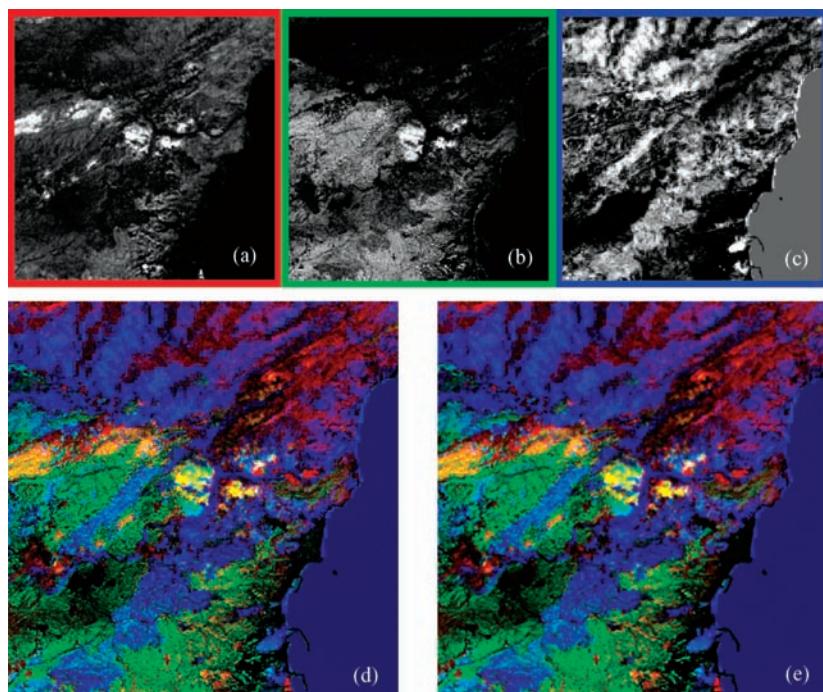
**Figure 20.10** (a) ASTER band 4 – band 6 differencing image; (b) ASTER band 4/band 6 ratio image; (c) spectral profiles of original ASTER bands 1–9 (VNIR and SWIR); and (d) spectral profiles of the BCET stretched ASTER bands 1–14 (VNIR, SWIR and TIR)

- Gypsum: Similar to the above two groups in SWIR bands, but there is a reflectance peak rather than absorption trough in NIR band 3. Recall the last case study, in which this spectral profile is similar to the ETM+ spectral profile of gypsum illustrated in Figure 20.3.
- Blue: band 4 – band 6 + band 11 – band 3. The difference between band 11 and band 3 eliminates argillic alteration for its low thermal emission in band 11 and suppresses gypsum for its high reflectance in band 3, leaving silica to be further enhanced as the brightest pixels, Figure 20.11c.

Based on the above spectral signatures, the following compound differencing colour composite can enhance the three different minerals in distinctive colours:

- Red: band 4 – band 6. Generally highlights all the alteration clay minerals and hydrated minerals, Figure 20.11a.
- Green:  $2 \times$  band 4 – band 6 – band 3. Because of the high reflectance of gypsum and strong absorption of alteration clay minerals in band 3, this difference eliminates gypsum while further enhancing alteration clay minerals, Figure 20.11b.

Therefore, in this compound difference colour composite (Figure 20.11d), gypsum is bright orange to reddish because it is only bright in the red layer; argillic alteration is bright yellowish as it is bright in both red and green layers; and quartz siliceous alteration is white as it is bright in all the RGB layers. There are many noise-like blocky edge effects in this image as the result of different spatial resolutions of the VNIR, SWIR and TIR band groups. This artefact can be effectively suppressed by a  $3 \times 3$  smoothing filter (Figure 20.11e).



**Figure 20.11** Generation of compound differencing colour composite: (a) band 4 – band 6 generally highlights clay and hydrated minerals; (b)  $2 \times$  band 4 – band 6 – band 3 eliminates gypsum and highlights clay minerals of the alteration zone only; (c) band 4 – band 6 + band 11 – band 3 highlights quartz; (d) compound colour composite of (a) in red, (b) in green and (c) in blue; and (e) compound colour composite smoothed with a  $3 \times 3$  smoothing filter

Regional prospecting using the ASTER image successfully located an alteration zone and separated two different types of alterations. However, for more detailed study of these detected small alteration targets, the relatively low spatial resolution of the ASTER image is obviously not adequate. Thus an airborne remote sensing study at high spatial resolution focusing on this alteration zone is required.

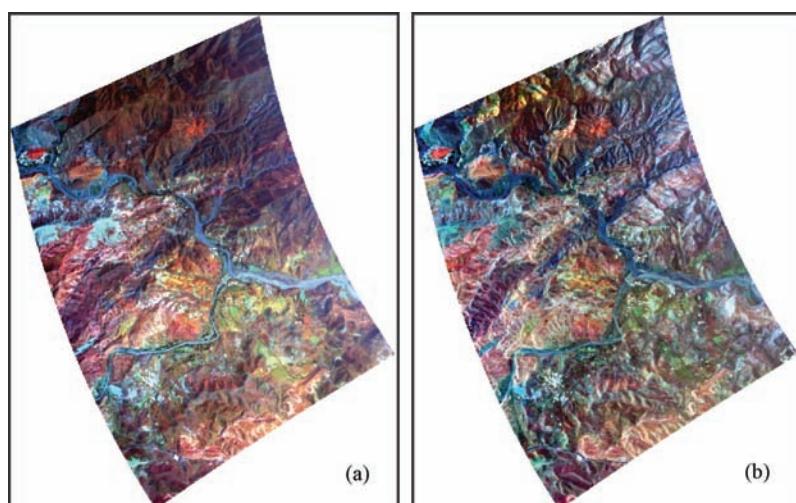
### 20.2.3 ATM image processing and analysis for target extraction

The ATM is an airborne version of the Landsat TM/ETM+ across-track scanner but it has finer spectral resolution with a total of 11 spectral bands (10 reflective spectral bands in the VNIR and SWIR spectral regions and 1 broad thermal band) as listed in Table 20.3. The spatial resolution of the ATM is dictated by the flight altitude, which is 7.5 m for this dataset.

With the regional prospecting results using the ASTER image, the obvious starting point of the detailed study using the ATM image is to repeat the technique used to produce a higher resolution version of the equivalent results generated from the ASTER image. One of the advantages of the ATM image is that its thermal band has the same spatial

resolution as the reflective spectral bands and therefore it is possible to generate high-quality colour composites of simulated reflectance as introduced in Section 3.7.

We start from simple colour composites and simulated reflectance colour composites. Figure 20.12 shows colour composites of (a) ATM bands 10–5–2 in RGB with DDS enhancement and (b) simulated reflectance of the same bands. This band combination is equivalent to TM bands 731 in RGB displaying the clay mineral absorption SWIR band in red, the red band in green and the blue band in blue. The DDS and the simulated reflectance colour composite are spectrally similar but the topographic shadows in the simulated reflectance colour composite are subdued with spectral variation further enhanced. Comparing these images with the ASTER compound difference colour composite in Figure 20.11, the cyan patches on the left of the image are gypsum outcrops and the brown-coloured patch left of the river junction where two channels merge into one is the epithermal alteration zone. The much higher spatial resolution of the ATM data indeed brings out many details of these mineral targets. Among others, the most eye-catching features in these two images are the red patches in the top half of the images. These are iron ore deposits in ferroan dolomites. Showing in red, these iron deposits are characterized by high



**Figure 20.12** ATM bands 10–5–2 colour composites: (a) DDS enhanced; and (b) simulated reflectance

reflectance in ATM band 10 where both alteration clay minerals and gypsum have strong absorption, and therefore are not depicted by the ASTER compound differencing colour composite in Figure 20.11.

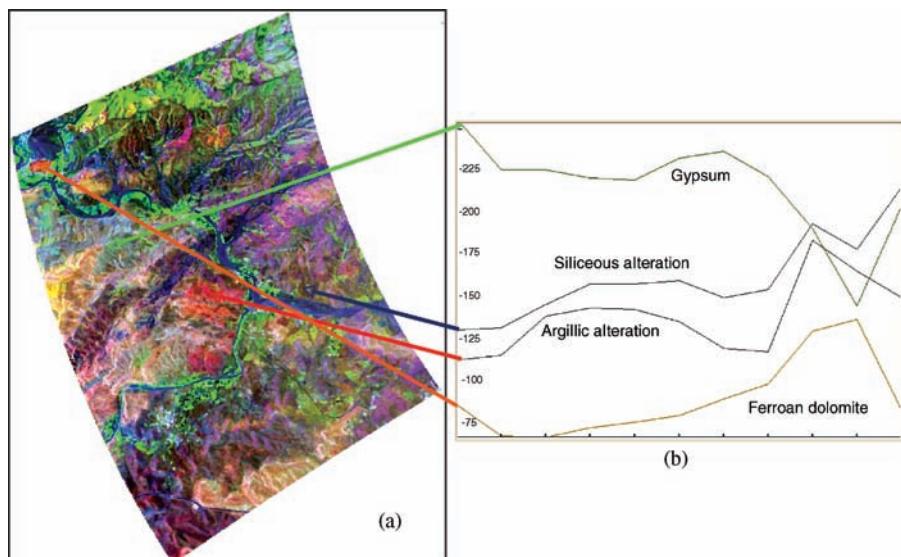
The simulated reflectance colour composite of ATM bands 972 in RGB (Figure 20.13a) displays similar phenomena as the RGB composites of bands 10–5–2 except for vegetation in distinctive bright green, but the iron deposits are less red and the epithermal alteration zone is more reddish, making the two less distinguishable. The spectral profiles extracted from the BCET-processed ATM image data (Figure 20.13b) indicate that the ferroan dolomite has high reflectance in both ATM bands 9 and 10 but the reflection peak is at band 10, which is a diagnostic feature distinguishing ferroan dolomite from all the other mineral targets in the study area. The ATM spectral profiles of argillic alteration, siliceous alteration and gypsum are similar to those obtained from the ASTER image shown in Figure 20.10d. As a trial, the compound difference colour composite using an ATM band combination equivalent to the ASTER compound differencing colour composite in Figure 20.11d was produced as follows:

- Red:  $ATM9 - ATM10$  highlights both the argillic alteration and gypsum.

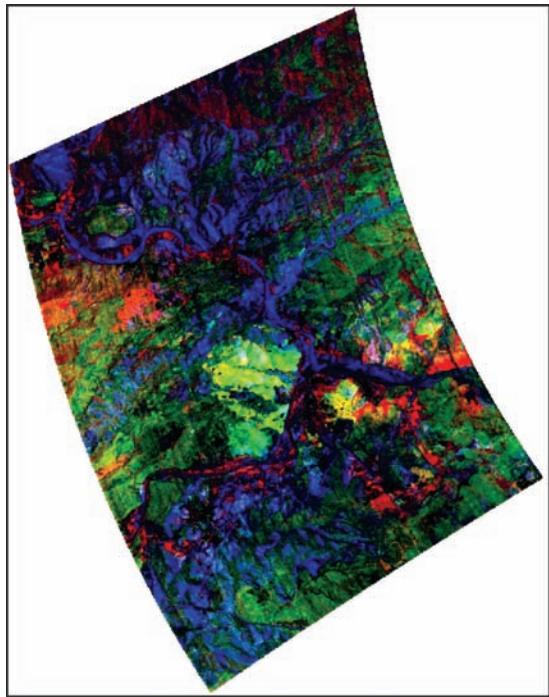
- Green:  $2 \times ATM9 - ATM10 - ATM8$  highlights argillic alteration only.
- Blue:  $ATM9 - ATM10 + ATM11 - ATM7$  highlights siliceous alteration.

Following this same principle, the image in Figure 20.14 presents argillic alteration in bright yellow to green, gypsum in bright orange red and siliceous alteration in white. However, the ATM band 10 is a rather broad SWIR band in comparison with the ASTER band 6 which targets the deepest absorption of most clay and hydrated minerals. The same is true for the thermal band, which depicts the diagnostic thermal emission feature of quartz, whereas the ATM compound difference colour composite does not enhance these minerals as distinctively as the ASTER one. Unsurprisingly, the image does not enhance the iron deposits in ferroan dolomites either, since it is not designed for the purpose.

As mentioned before, the key spectral feature making the iron ore deposits in ferroan dolomites different from the argillic alteration zone and gypsum is the higher reflectance in ATM band 10 than in band 9, but this feature is shared by many other rock types and is thus not diagnostic. We would like to produce a colour composite that enhances ferroan dolomite, the argillic alteration zone and gypsum only. Since this is not easily achieved using



**Figure 20.13** (a) Colour composites of simulated reflectance of ATM bands 972 in RGB; and (b) spectral profiles of gypsum, siliceous alteration, argillic alteration and ferroan dolomite derived from BCET stretched ATM image data



**Figure 20.14** The compound difference colour composite of  $ATM9 - ATM10$  in red,  $2 \times ATM9 - ATM10 - ATM8$  in green and band 4 – band 6 + band 11 – band 3 in blue

arithmetic operations, we consider a combined approach using both differencing and the feature-oriented PC selection method (FPCS, see Section 7.2). PCA was applied to bands 2 to 10; band 1 of this

ATM dataset is very noisy and was therefore discarded and band 11, the thermal band, was also discarded as it is not relevant for the target features. Table 20.4 presents the matrix of eigenvectors of the covariance matrix of the nine bands. PC5 is dominated by the difference of ATM band 10 and band 9 depicting the key spectral feature of ferroan dolomites. The difference between band 7 and band 5 enhances vegetation as well in PC5. As shown in Figure 20.15a, ferroan dolomites are bright while both argillic alteration and gypsum are very dark in the PC5 image. PC3 is largely a weighted summation of all the VNIR bands subtracting the summation of SWIR bands 9 and 10. As implied in the ATM spectral profiles in Figure 20.13b, the operations for PC3 will produce high values for gypsum and very low values for argillic alteration zones and the ferroan dolomites, and thus the negative PC3 highlights argillic alteration zones and ferroan dolomites while suppressing gypsum as shown in Figure 20.15b. Again we use the difference image of  $2 \times ATM9 - ATM10 - ATM8$  to highlight argillic alteration only, as shown in Figure 20.15c. Finally a colour composite is generated as below:

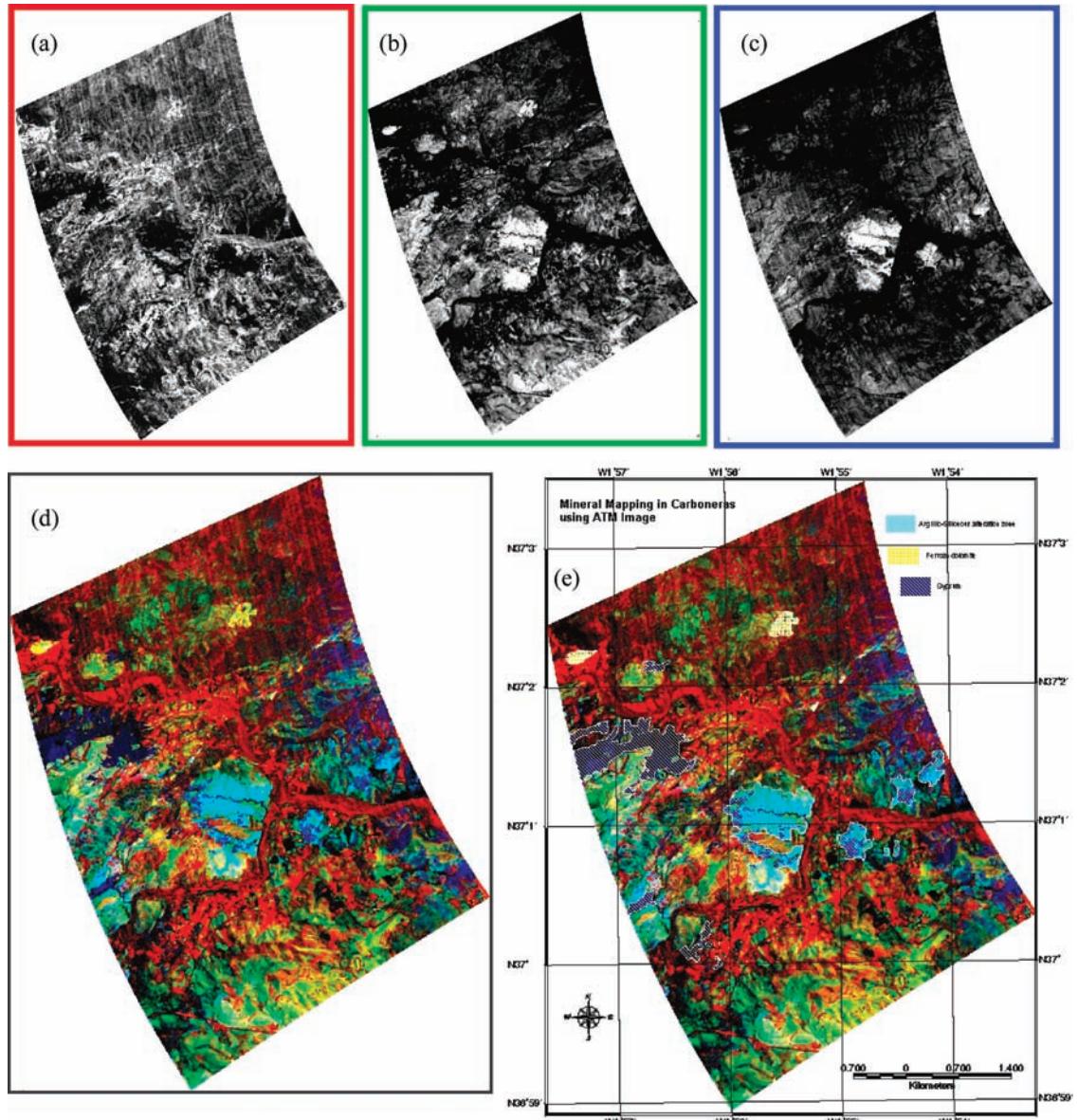
- Red: PC5 highlights ferroan dolomites and suppresses argillic alteration and gypsum.
- Green: Negative PC3 highlights both ferroan dolomites and argillic alteration and suppresses gypsum.
- Blue:  $2 \times ATM9 - ATM10 - ATM8$  highlights argillic alteration only.

**Table 20.4** Eigenvectors of the covariance matrix of 9 ATM image bands.

Covariance Eigenvectors	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
Band 2	0.320	0.395	0.292	-0.679	0.196	-0.046	0.354	-0.160	0.065
Band 3	0.348	0.325	0.187	-0.046	-0.019	0.054	-0.631	0.450	-0.363
Band 4	0.343	0.286	0.091	0.320	-0.215	-0.162	-0.184	-0.117	0.755
Band 5	0.339	0.199	0.071	0.430	-0.297	-0.192	0.298	-0.395	-0.535
Band 6	0.347	-0.208	0.234	0.262	0.039	0.481	0.476	0.495	0.089
Band 7	0.331	-0.490	0.248	-0.038	0.223	0.342	-0.353	-0.544	-0.012
Band 8	0.312	-0.558	0.078	-0.154	-0.064	-0.704	0.035	0.247	0.002
Band 9	0.329	-0.101	-0.628	-0.347	-0.539	0.276	-0.008	-0.013	0.013
Band 10	0.329	0.111	-0.592	0.188	0.693	-0.106	0.036	0.013	-0.014

The resulting colour composite in Figure 20.15d shows ferroan dolomites in yellow because they are bright in both red and green layers, and argillic alteration in cyan because of its high values in green and blue layers. It is interesting to notice

that the image depicts gypsum in very distinctive deep blue, while none of the three images forming this colour composite highlights gypsum. As shown in Figure 20.15, gypsum is suppressed as very dark features (this is distinctive as well!)



**Figure 20.15** (a) PC5 image of ATM bands 2–10; (b) negative PC3 image of ATM bands 2–10; (c) compound difference image of  $2 \times \text{ATM9} - \text{ATM10} - \text{ATM8}$ ; (d) colour composite of (a), (b) and (c) in RGB; and (e) interpretation of the three major mineralization targets in the area: argillic–siliceous alteration zone, ferroan dolomite and gypsum

in both PC5 and negative PC3, while in the compound difference image it is in medium grey. Consequently, gypsum is clearly enhanced as deep blue in the resulting colour composite (Figure 20.15d). This reminds us that feature enhancement can be achieved by suppression as well, instead of by highlighting!

Finally, as shown in Figure 20.15e, a simple interpretation map of the argillic alteration zone, iron ore deposits in ferroan dolomites and gypsum outcrops was produced from the PC and compound difference colour composite image.

#### 20.2.4 Summary

In this case study, we used ASTER and ATM images to demonstrate the application of multi-spectral and multi-resolution remote sensing for mineral exploration via image processing. Firstly, the ASTER image with lower spatial resolution but large coverage was processed and analysed for regional prospecting. This enabled us to focus on a much smaller area with several different types of minerals, using a higher spatial resolution ATM image for detailed study. For both datasets, image processing began from optimal visualization of the data followed by well-established standard enhancement techniques for the presumed targets. Then image spectral profiles of pixels representing possible mineral targets located by standard techniques were carefully analysed to design further processing strategy and specific enhancement operations addressing the diagnostic spectral features of the target minerals. Using compound difference images and the FPCS PCA method, regional distributions of argillic-siliceous alteration and gypsum are highlighted in an ASTER compound difference colour composite image, while the details of argillic–siliceous alteration, iron deposits in ferroan dolomites and gypsum are mapped by a colour composite of PC and compound difference of an ATM image. One interesting lesson to learn from this ATM colour composite in its enhancement of gypsum is that enhancement does not always mean highlighting the target in bright pixels: suppressing the target features as very dark pixels is enhancement as well!

### 20.3 Remote sensing and GIS: evaluating vegetation and land-use change in the Nijar Basin, SE Spain

#### 20.3.1 Introduction

This case involves the use of multi-temporal satellite image datasets acquired during the period between 1984 and 2004, to demonstrate the nature, distribution and rate of change to land-use patterns in the Nijar Basin, in the Almeria Province of south-east Spain. It is a fairly simple case which demands accurate data co-registration (georeferencing) and the processing of multi-spectral imagery to reveal features characteristic of land use in this area, in order to identify where and how much change has occurred. GIS here serves to manage and display the processed results and to enable some spatial statistical analysis.

The sustainable economy of Almeria Province has for some considerable time been based on agriculture, and this is still the case today. Throughout the 1980s the style of agriculture changed radically from one of open growth of grapes, olives, nuts and other vegetables to the highly intensive production of tomatoes, melons, cucumbers, strawberries and other soft fruits and vegetables, under plastic in greenhouses. Flat ground, plentiful sunshine and EU subsidies have together enabled the rapid development of this style of agriculture. The plastic covering prevents excessive evaporation of water, helps keep pests out and promotes a year-round growing season. This growth has been accompanied by a huge increase in demand for water, which has traditionally been supplied by an aquifer located in the Messinian sediments below. Unregulated pumping on an unprecedented scale eventually caused a depression in the local groundwater table and the incursion of saline water from the Mediterranean Sea to the south. These events have been the subject of some attention and many publications exist. Irrigation styles have now changed to drip-feed methods which use water much more effectively.

An air flight to Almeria these days greets the tourist with a view of the ‘sea of plastic’ which now covers the much of the open, flat ground around Almeria. As the aircraft comes into land, it is a

shocking and spectacular sight. The construction of such plastic greenhouses is today a very hot political topic since the benefit to the economy is undeniable yet most agree that they are a ‘blot on the landscape’.

The location of the Nijar Basin with respect to the other previously described teaching case study areas is shown in Figure 19.1. We have already described the climatic setting of this area, in Section 19.3.3, as being semi-arid (semi-desert). The 200–300 mm incident rainfall predominantly falls on the Sierras, in the months between October and April, and of that rainfall perhaps 40–50% is lost through evaporation and 10–20% lost as runoff so that perhaps 30% infiltrates and becomes groundwater and ultimately enters the aquifer.

The Nijar Basin lies between the Sierra de Alhamilla and a range of low hills, called La Serrata, which represents the topographic expression of the Carboneras fault. The latter forms a structural trap into which terrestrial sediments have been deposited. The resulting sequence of marls, gypsum, limestones and sandstones deposited in the basin now

forms the main aquifer supplying water for agriculture in the Nijar area. The geological setting is illustrated by the interpretation map shown in Figure 19.7 and the local geography is shown in Figure 20.16.

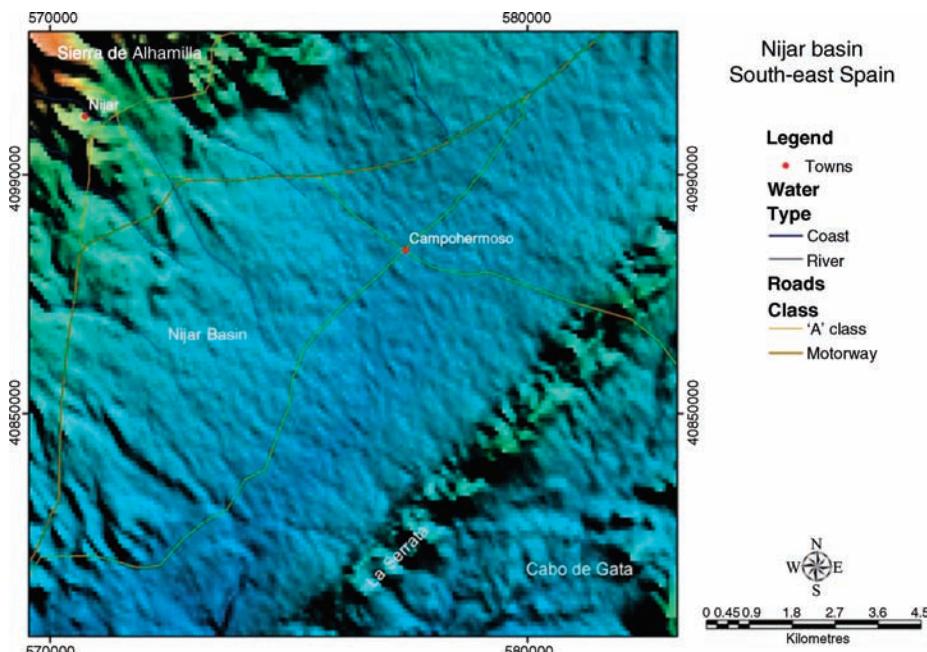
The objectives of this study are broadly to demonstrate the use of multi-temporal imagery in revealing land-use change in the Nijar Basin. In doing so, we hope to comment on the rate and nature of the changes, and to illustrate these graphically and quantitatively.

More specifically, the objectives are to:

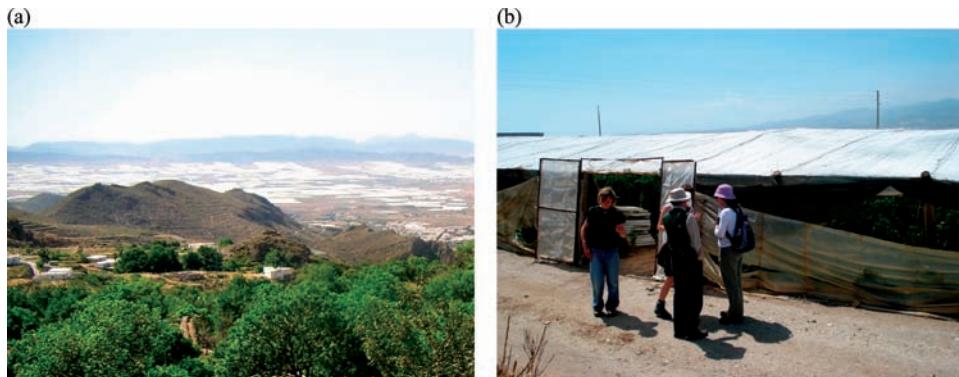
- Highlight the distribution of both vegetation (natural and agricultural) and plasticulture (see Figure 20.17) for each year represented in our database.
- Locate and quantify areas of changing land cover.

### 20.3.2 Data preparation

In any study of temporal change, a set of image data acquired over a time period is required. In this



**Figure 20.16** Map of the Nijar Basin area with SRTM shaded relief as background topography. The field of view and scale bar here relate to all the images shown in this case



**Figure 20.17** (a) View looking south-eastwards, from the lush, spring-fed vegetation of the upland village of Huebro, over the greenhouses of the Nijar Basin in the middle distance; and (b) one of the older plastic-covered greenhouses constructed from wooden posts and covered with plastic-coated fabric mesh

particular case, the area in question has undergone quite a radical change over a period of 20 years and so the database consists of subsets of four Landsat scenes, acquired in 1984, 1989, 1992 and 2000, and digital aerial photographs acquired in 2004, as summarized in Table 20.5.

Generally speaking, these days all Earth observation (EO) datasets are delivered as georeferenced products. In the 1980s and early 1990s, however, this was not the case, so the first step here is to georeference, or at least co-register, the older image datasets used, and to ensure that this is done as accurately as possible. The Landsat-7 ETM+ dataset already conforms to WGS84 and UTM (zone 30), so the most logical step is to co-register the raw Landsat-5 images to the Landsat-7 ETM+ scene. As explained earlier, given that all the images were acquired by similar sensors and have similar geometry, this co-registration is best achieved using a simple linear translation and rotation. As explained in 19.1.2, BECT has been applied to all datasets.

### 20.3.3 Highlighting vegetation

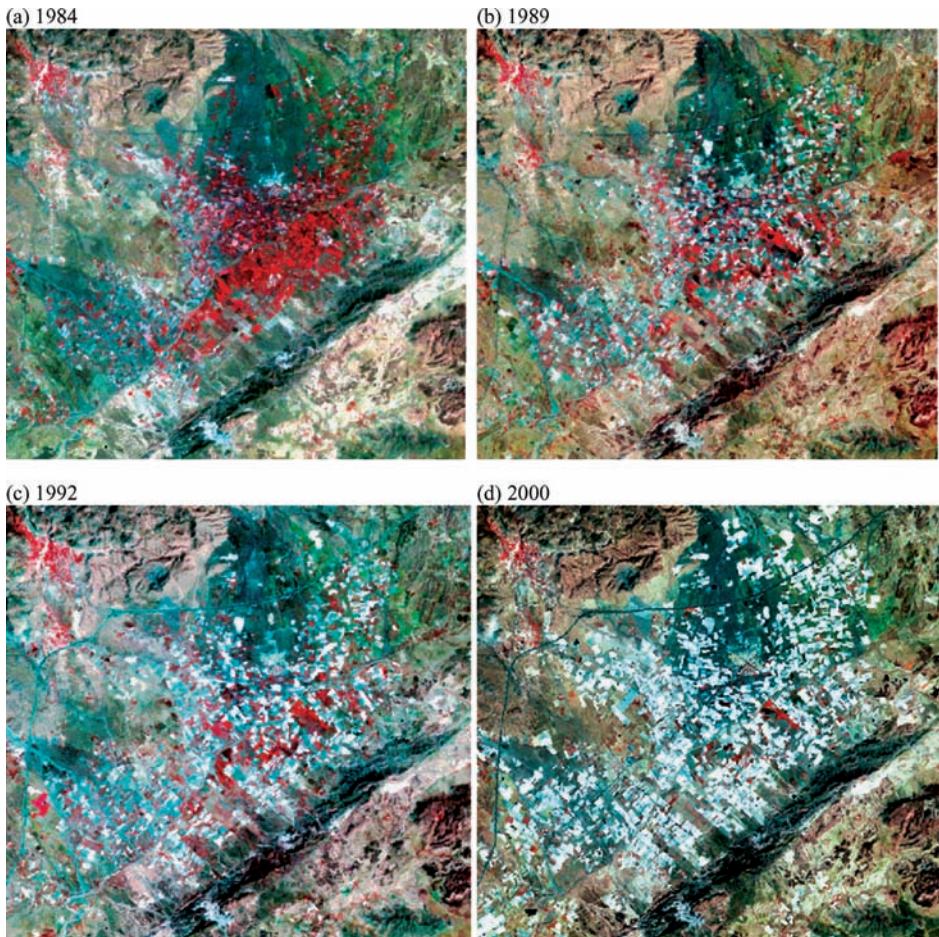
Here we begin with standard false colour composite images, of bands 432 RGB, for general visualization of land cover types (vegetated and non-vegetated) for each of the years of observation. For all four years, these are shown in Figure 20.18.

We can see clearly from the extent and form of the dense patch of red in 1984 that the vast majority of land under agriculture was not devoted to plasticulture but was under open skies (Figure 20.18a). We can also see that plasticulture made its first impact in 1989 and expanded steadily to 2000.

Now that we have identified the vegetated and plasticulture areas we need to establish both the decrease in open agriculture and the expansion of plasticulture (since they do not necessarily mirror one another). Firstly, we need to identify and extract the open vegetation so we will use a simple normalized difference vegetation index or NDVI. The resultant NDVI images for each year are shown in Figure 20.19. These allow us to visualize the extent

**Table 20.5** Multi-temporal datasets used in the case study

Dataset	Scene identifying numbers	Acquisition date(s)
Landsat-5 TM	Subset of path 199/row 034	1984, 1989, 1992
Landsat-7 ETM+	Subset of path 199/row 034	2000
Aerial photography	Ortho-quads (Junta de Andalucia air survey)	2004

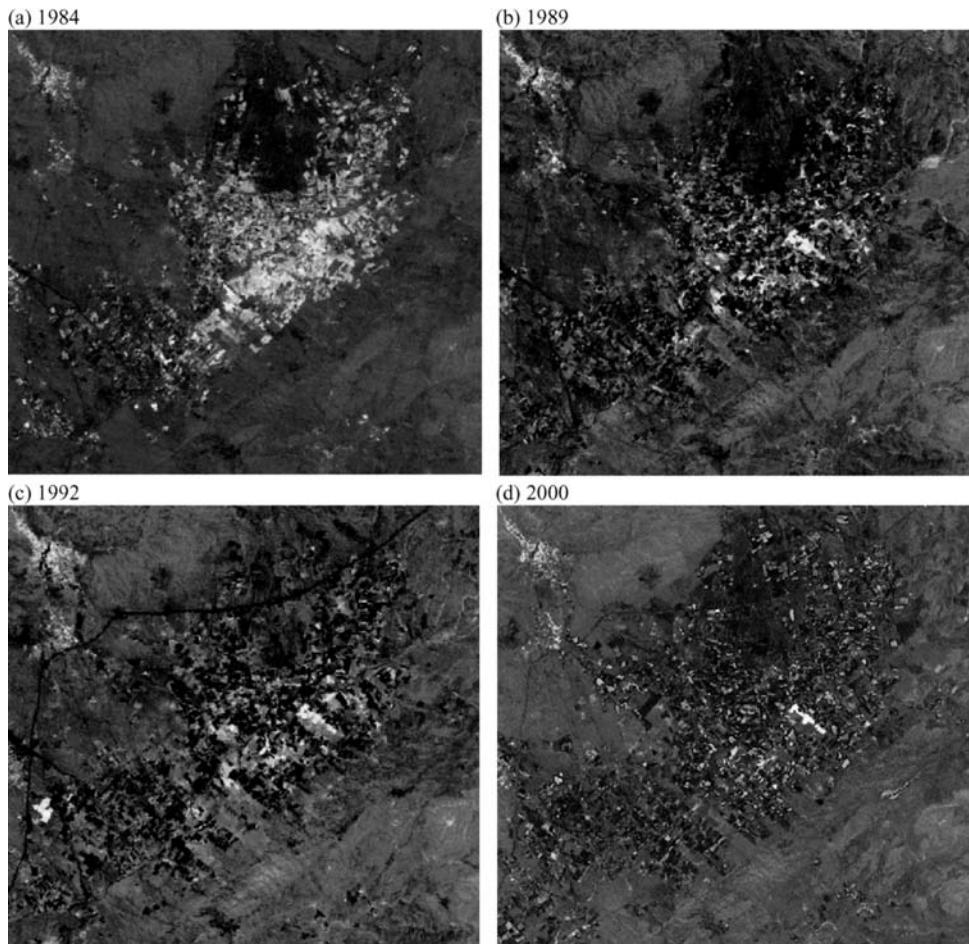


**Figure 20.18** Standard false colour composite images (432 RGB) illustrating clearly the extent of open vegetation (in red tones) and plastic greenhouses (in white and bright cyan tones) as observed in: (a) 1984; (b) 1989; (c) 1992; and (d) 2000. Rocks and soils, exposed beyond the cultivated and urban areas, appear in a variety of grey, bluish green, brown and greenish tones

and decline of open vegetation from 1984 when almost no plasticulture existed (Figure 20.19a), which we take as the baseline for our estimations, to 2000 when very little open agriculture remained. In 2000, the few patches of healthy vegetation represent gardens and trees growing around the town of Níjar, which is supplied with plentiful spring water, and a few isolated fields and old greenhouses which are in disrepair. One large patch of open agriculture is noticeable (centre right) in the NDVI of every year; this represents a farm whose owner stubbornly refuses to adopt plasticulture. Sadly, this plot is now in a state of disrepair.

Surrounding the central patch of cultivation in each image, the background appears as a mid-grey tone; this represents the natural scrub vegetation characteristic of this semi-arid part of Spain and has a fairly similar appearance in every year.

The 1992 image (Figure 20.19c) contains a linear feature running from the bottom left to top right, which has anomalously low NDVI values. This represents the main Malaga–Murcia motorway, which was constructed around 1991–1992. The motorway as shown in 2000 is represented by a much narrower dark line than in 1992. This is probably a result of the re-establishment of natural



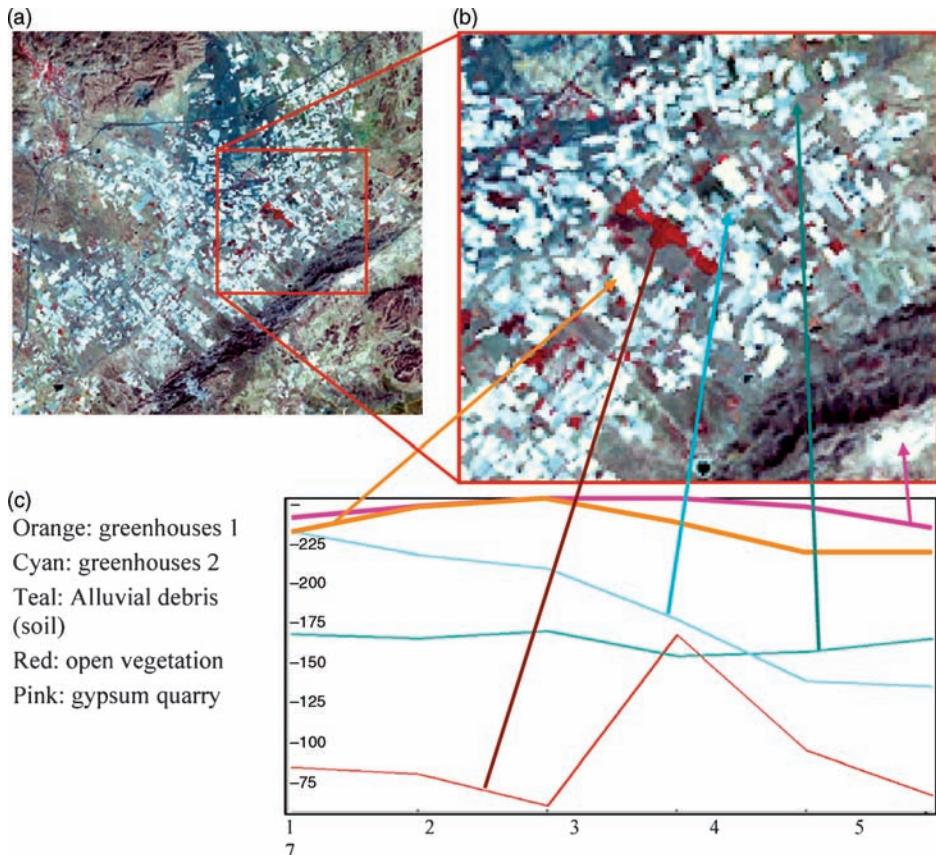
**Figure 20.19** NDVI as calculated from each image dataset: (a) 1984; (b) 1989; (c) 1992; and (d) 2000

vegetation on the verges of the motorway, as opposed to the broad swath of ground which was stripped of vegetation during construction in 1992. Conversely, the area around Nijar town is relatively bright in each year, since the vegetation here is found in parks and gardens inside and around the town.

#### 20.3.4 Highlighting plastic greenhouses

To highlight and extract the plastic-covered greenhouses we must first understand their spectral properties. The roofs are what we see in these images and, on occasion, we may see something

of the vegetation growing within. These roofs are generally highly reflective and often cause saturation in the visible bands. The roofs are sometimes painted dark (black or grey) in winter to make them less reflective and thus absorb more radiation. In some cases where the plastic sheeting is relatively new and/or less opaque than older greenhouses, or where the sheeting has not yet been painted, chlorophyll in the growing vegetation inside makes a contribution to the overall reflectance of the greenhouse and it appears pale pink in the standard false colour images shown in Figure 20.18. In other cases, the plastic roofs appear less reflective because they are curved rather than flat (generally the more modern constructions).



**Figure 20.20** (a) True colour composite (321, 2000) of the Nijar Basin area; (b) detail of the image in (a); and (c) Landsat ETM+ spectral profiles of some of the main ground targets for discrimination

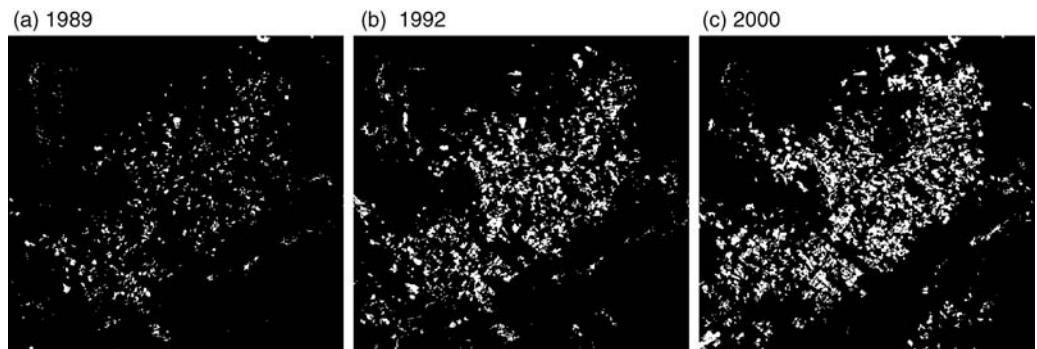
The plastic itself, though, does not appear to have any particularly diagnostic spectral features and so its signature resembles that of other highly reflective targets, such as bare, flat, cleared fields or the smooth level floor of a nearby gypsum quarry. Clearly, the plastic-covered targets in the images cannot simply be extracted on the basis of visible brightness alone. As with any other target, we need to look carefully at the spectral signatures to discover some diagnostic features and thus a way to separate them from the other image features. Using one of the images in the database for illustrative purposes (Landsat ETM+ 2000 in this case), the spectral profiles of some of the main ground targets are shown in Figure 20.20.

With these observations of spectral profiles in mind, a simple formula is designed to extract pixels

representing plastic-covered greenhouses using the following algebraic and logical operations:

$$\begin{aligned}
 &\text{If } ((b4 - b3)/(b4 + b3)) < 0 \text{ and } (b5/b7) < 1.1 \\
 &\quad \text{then (if } ((b1 + b2 + b3)/3) > 225 \\
 &\quad \quad \text{then 1 else 0) else 0.} \quad (20.3)
 \end{aligned}$$

Referring also to the profiles in Figure 20.20, the first part  $((b4 - b3)/(b4 + b3)) < 0$  effectively removes any surface which is vegetated (cultivated or natural), the second part  $((b5/b7) < 1.1)$  enhances and thresholds hydrated minerals (including gypsum), and the last part  $((b1 + b2 + b3)/3) > 225$  masks on the basis of average visible brightness. All excluded pixels are then coded as 0 and all retained pixels coded with a value of 1, thus producing the binary images shown in Figure 20.21.



**Figure 20.21** Binary images with pixels encoded to show the extent of plasticulture for the years 1989, 1992 and 2000 (we have assumed that there were no greenhouses in 1984)

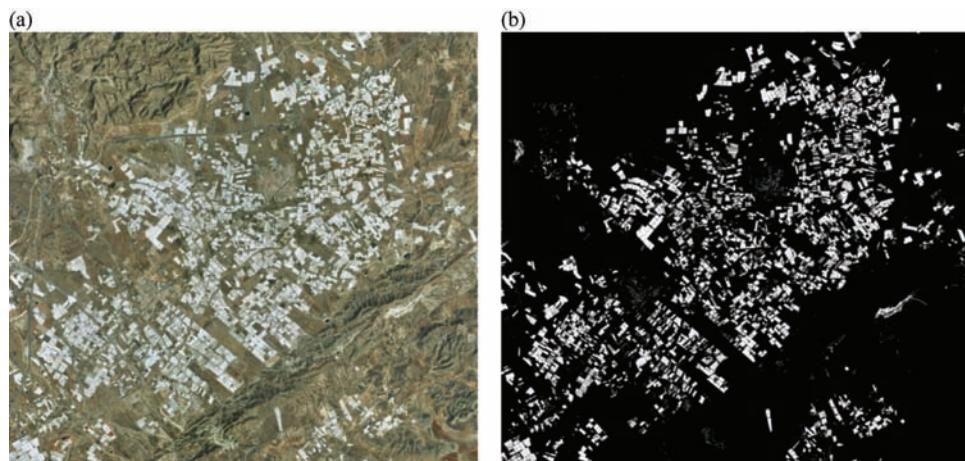
The differing illumination conditions on each acquisition date mean that the thresholds in (20.3) need to be adjusted to compensate for slight changes in relative brightness in each image. Even after such minor adjustments, we find that in Figure 20.21c we have gained some pixels from the road leading to a small gypsum quarry and have partially lost one large, almost triangular-shaped greenhouse from the upper centre of the area. Our algorithm excluded the latter because its roof had been painted sometime before 2000 and is much darker in 2000 than it was in either 1989 or 1992.

Our dataset also contains ortho-rectified digital aerial photographs (as a three-band true colour image only, i.e. without near infrared) acquired in 2004. These images have 1 m spatial resolution and have been mosaiced to produce the image shown in

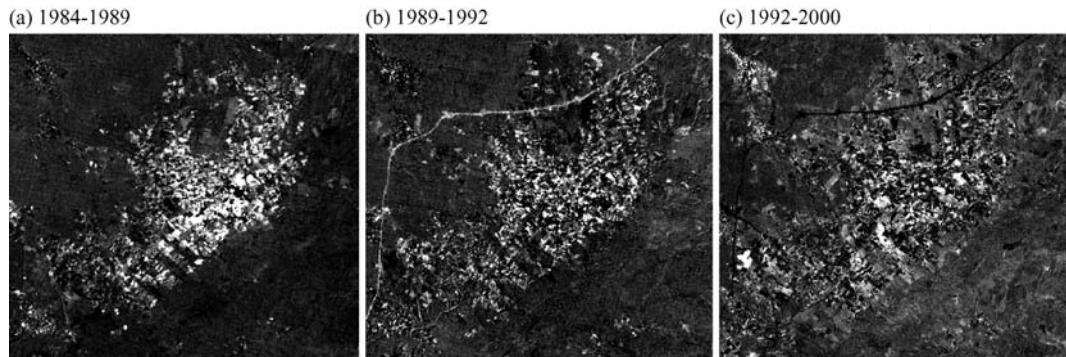
Figure 20.22a. They provide high spatial detail for the interpretation of ground features, including plastic-covered greenhouses. If we perform a similar procedure to classify the greenhouses from this image, though this time on the basis of relative visible brightness since we have no infrared bands, we produce the binary image in Figure 20.22b. Once again we have picked up the few pixels that make up the small gypsum quarry, but we will ignore these.

### 20.3.5 Identifying change between different dates of observation

Since the images are now accurately co-registered, we can used image subtraction to identify areas of change from one image to another. We may also



**Figure 20.22** (a) Ortho-photo true colour mosaic (2004) of the study area; and (b) the binary plasticulture image derived from it



**Figure 20.23** Difference intervals between NDVI images: (a) 1984–1989; (b) 1989–1992; and (c) 1992–2000. High values (bright pixels) indicate a loss of vegetation in the time period

display different dates in different colour guns. Firstly, we need to consider exactly what we need to compare in this way, so we should remember two of the primary objectives: highlight areas of vegetative cover and highlight plastic greenhouses. With these in mind, we can use the NDVI as one measure from which to identify change and the thresholded plasticulture image produced in Figure 20.23 as the other. We can then compare these indices for each year to estimate the proportion of land devoted to open cultivation.

The difference images in Figure 20.23 indicate that the greatest reduction in the extent of open vegetation occurred between 1984 and 1989. The white line that appears in Figure 20.23b represents the loss of vegetation caused by motorway construction in 1991–1992; the same motorway path in Figure 20.23c shows the opposite change in this time period and so appears black, representing the re-establishment of vegetation adjacent to the motorway. The difference image Figure 20.23c also shows a bright patch around the town of Níjar, representing a decrease in open vegetation between 1992 and 2000; the reasons for this are unclear and would require field investigation to explain.

Thresholding of the NDVI selectively to retain the highest values representing healthy cultivated vegetation, reclassifying them to a value of 1, and excluding the remaining values representing natural scrub vegetation and unvegetated areas (classified to a value of 0), allows us to produce a binary image representing open cultivation. If we perform this classification for each year, we may calculate the proportion of the area devoted to open cultivation and therefore gain an indication of the reduction over the time period, using any GIS statistical package. Table 20.6 shows the results of such calculations; we find that land devoted to open cultivation has decreased from a maximum of 8% in 1984 to a minimum of less than 1% in 2000.

Using the thresholded binary images shown in Figure 20.21 in which we have classified plasticulture with a value of 1 against all other pixels with a value of 0, we can again calculate the areas and proportions, this time of land occupied by plasticulture. This reveals that from our baseline in 1984, when we observe only open cultivation, plasticulture has commenced and increased to the point where it occupies more than 10% of the total study area, as summarized in Table 20.7. In other areas,

**Table 20.6** Statistics of change: land devoted to open cultivation

Year	1984	1989	1992	2000
Open cultivation ( $\text{km}^2$ )	11.7	3.2	3.5	1.3
Non-cultivated or plasticulture ( $\text{km}^2$ )	155	164	164	166
Proportion of the area	8%	2%	2%	<1%

**Table 20.7** Statistics of change: land devoted to plasticulture

Year	1984	1989	1992	2000	2004
Plasticulture area (km <sup>2</sup> )	<1	2.8	5.1	15.6	20
Non-plasticulture area (km <sup>2</sup> )	>166	164.2	161.9	151.4	148
Proportion of the area	<2%	2%	3%	10%	>13%

such as south-west of Almeria city, this percentage is far higher. Making the same calculation of area from this binary image, we find that the proportion of land devoted to plasticulture has increased again to just over 13%. These figures seem rather low in visual comparison with the images but they do not account for the averaging that the human eye/brain tends to perform, which ignores the areas in between the greenhouses; some adjustment would be required to account for this to gain a more representative idea of land devoted to plasticulture.

### 20.3.6 Summary

This case study has allowed us to make an estimate of the extent of vegetation and plasticulture in this one small area, as being representative of change which is mirrored in other parts of the region. We have explained one method for doing so and some of the difficulties along the way. Clearly some assumptions and inaccuracies must be accepted, such as those introduced by the subjective application of thresholds to produce binary classifications, and these must always be recognized and acknowledged even if they cannot be eradicated.

Comparison of the extent of plasticulture in 2004 with that of open vegetation in 1984 reveals that plasticulture has more than merely replaced traditional agriculture, since it has expanded far beyond, even to some areas that might not at first seem suitable, such as the top of the La Serrata ridge.

It is tempting to try to classify the greenhouses themselves and thereby the type of vegetation growing inside them but this has, time and again, proved to be a waste of effort, for a variety of reasons. In many cases, there is indeed a contribution from the photosynthesizing plants to the overall reflectance of the greenhouse roof material but the amount of contribution is dependent on many

things, not least the age and type of material making up the roofs of the greenhouses. None of the roofs are of clear plastic or glass, some have been made more or less opaque though painting, some are made of mesh, and some have several layers of plastic or mesh, added over the years. Since greenhouses are costly to maintain, the only safe conclusion is that those with intact roofs are generally filled by some growing cash crop.

The availability of, and impact on, water resources in this region are further avenues for investigation but one in which remote sensing can contribute only to a limited degree since they occur underground. What we can surmise is that in this period (1984 to 2000) the changing style, extent and intensity of cultivation will have meant that the demand for water increased enormously. Unregulated pumping in the past from wells in this area has caused depletion of the local aquifer, which is known to have resulted in the incursion of saline water from the Mediterranean to the south, thus contaminating the aquifer and increasing salinity in soils. New irrigation techniques and regulation of water extraction have lessened these effects but water and soil quality remain important issues in this area and require constant monitoring.

## 20.4 Applied remote sensing and GIS: a combined interpretive tool for regional tectonics, drainage and water resources

### 20.4.1 Introduction

This study involves the use of multi-spectral imagery to improve our understanding of the

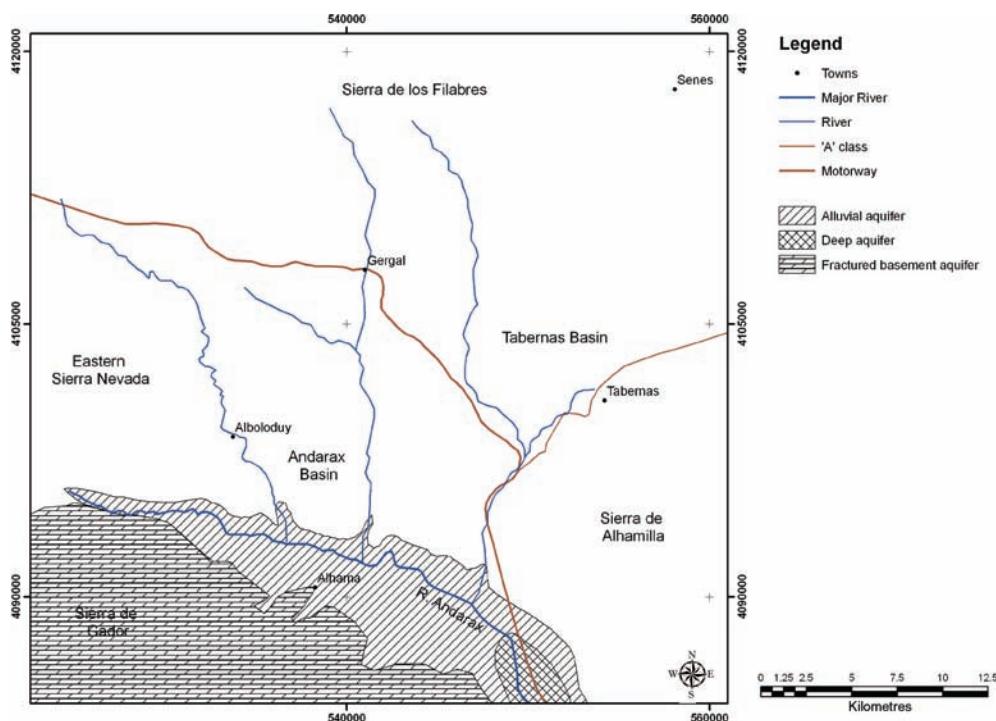
regional geology, tectonics and hydrology in the Tabernas–Andarax Basin of Almería Province, Spain. The focus is on the applied use of remote sensing and GIS in the interpretation and exploration of a region's water resources. While the topic seems to concern mainly geological concepts, the reason for concentrating on these phenomena is to reveal the controls on the sustainable rural economy in this semi-arid area. The geography and location of this case study area are illustrated in the map in Figure 20.24 and in Figure 19.1.

#### 20.4.2 Geological and hydrological setting

Unlike the Níjar Basin to the east, this area is predominantly devoted to open agriculture with the use of greenhouses only now increasing steadily. Water has been plentiful enough here to support the prolific growth of citrus fruits, not possible in the Níjar area, but only through systematic irrigation. The area under active irrigation is reported to have

increased to some 14 000 ha in the last 20 years (Pulido-Bosch *et al.*, 1994; Gallego *et al.*, 2006).

Water resources are recycled continuously, resulting in a concentration of solutes and a gradual degradation of water quality. This degradation has been severe enough, in the last few years, to cause a marked decrease in productivity and a decline in local economic terms. By contrast, the areas of the Andarax Valley nearer to Almería city have experienced these effects far less since their economic stability is lifted by the growth of Almería city which it owes partly to the successes of plasticulture in neighbouring areas. Research by Pulido-Bosch *et al.* (1994) involved the analysis of piezometric data (collected by the IGME) in boreholes between the towns of Almería and Gádor along the Andarax. All boreholes show a significant decrease in water levels between 1973 and 1987. Their work also showed that the salinity of groundwater increases south-eastwards, towards Almería. Discharge from the basement aquifer occurs mainly through springs and below ground level directly into the River



**Figure 20.24** Map of the main aquifers in this area: fractured karstified basement, deep sedimentary and alluvial. Modified after Pulido-Bosch *et al.* (1994)



**Figure 20.25** The Andarax river valley looking west, with the Sierra de Gador in the left background. The barren higher slopes appear in great contrast to the lush growth in the valley floor

Andarax. Recharge is through winter rain and snowfall on the Sierra de Gador (Figure 20.25), and to an extent on the Sierra Nevada to the north-west.

The characteristic lithologies of the area range in age from the Permo-Triassic to Quaternary. Those of the Sierra Gador basement massif comprise Alpujarride nappes of dolomites, schists and marbles. The Andarax Basin is filled by a sequence of sedimentary units of Neogene (Messinian and Pliocene) and Quaternary age. To the north, the Sierra de los Filabres is dominated by mica schists (in this area). There are three main aquifers which provide groundwater to the area. These are deep, fractured and karstified basement carbonates of the Sierra de Gador, unconsolidated Messinian sediments in the Andarax Valley, and, above these, shallow Pliocene and Quaternary alluvial and deltaic sediments which are exposed along the length of the valley (Pulido-Bosch *et al.*, 1994).

The entire region is tectonically active today. The most recently recorded deformation in this area is related to movement along the Carboneras fault and has produced faults, oriented approximately NNW–SSE, with vertical displacements of the order of 10 m. These can be seen in the Quaternary fan deposits at many localities of this region. The Sierra de Gador, Sierra de los Filabres and Andarax Basin

are separated by much older major basin-bounding faults, whose displacement history is complex, many of which have been reactivated in recent times and now provide pathways for water.

#### 20.4.3 Case study objectives

The overall objective of this study is to demonstrate the uses of multi-temporal, multi-spectral imagery in improving the understanding of geology and water resources, in a semi-arid area which is supported, dominantly, by a rural economy. The value of land here is relatively low, solar energy is plentiful and so the real limiting factor on the sustainable economy has always been the availability of water. The questions are, therefore, what controls the presence of water and what can we glean from the data we have about its presence?

With remote sensing we ‘see’ only the surface of the ground and therefore only the exposed rocks, soils and vegetation; from these we must extract (or infer) information relating to water resources at depth. The study area is quite large ( $25 \times 25 \text{ km}^2$ ) so we chose medium-resolution imagery for the task (in this case Landsat). Our approach begins by enhancing surface features which then enable us to

interpret sub-surface phenomena: vegetation type and distribution; geology and potential aquifers; and the surface expressions of structural features, i.e. faults and joints (fractures in general), which act as pathways for water once it penetrates the surface.

The main objectives are therefore to:

- Locate natural and cultivated vegetation.
- Distinguish the main litho-tectonic units.
- Enhance and identify the main structural elements.
- Examine how these pieces of evidence reveal and explain the connections between land use, water and geology.

These goals will be achieved by:

1. Preparing a multi-temporal, multi-spectral, medium-resolution dataset.
2. Simple directed processing of multi-spectral imagery.
3. Interpretation of land cover, geology and geomorphology.

Our database for this study consists of multi-temporal Landsat images, digital aerial photographs, images and DEMs (both ASTER and SRTM) as summarized in Table 20.8. As in the previous example, all data conform to WGS84 data and UTM zone 30.

#### 20.4.4 Land use and vegetation

To get a first glimpse of the importance of water in this environment, one need only look at the distri-

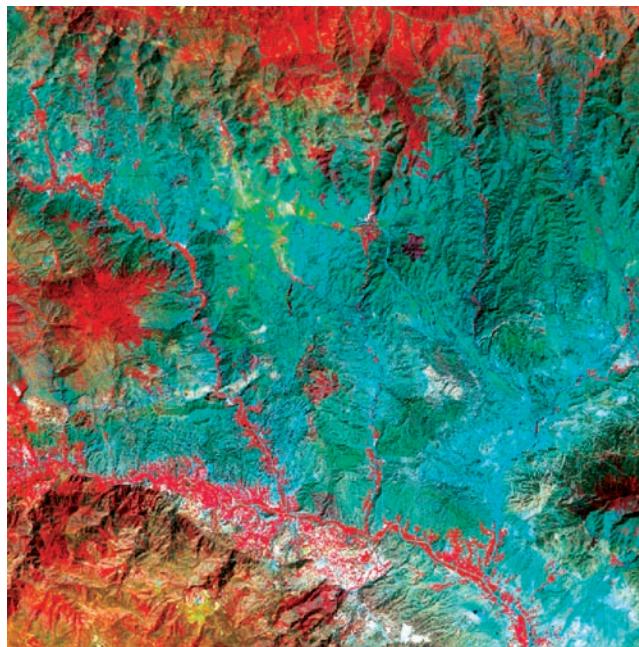
bution of vegetation since it cannot survive without water. Both the cultivated vegetation and natural plant cover are important here. The transportation of water is expensive and might be prohibitive to economic agriculture if it were necessary, so the existence of extensive areas of cultivated land suggests that plentiful, natural water supplies are in close proximity. The density and intensity of cultivation along the Andarax river valley suggests that water is indeed plentiful. In contrast, the distribution of natural vegetation on adjacent hillsides reveals several things, including the whereabouts of the main recharge areas (where rainfall is incident), areas where the rocks and soils hold water and those where evapotranspiration is relatively low (largely on north-facing slopes).

Standard false colour composites of bands 432 (RGB) reveal significant patterns in the distribution of vegetation (Figure 20.26). The most obvious one is the marked absence of vegetation in the centre of the area, in the ‘badland’ terrain occupied by Messinian and Pliocene sediments of the Tabernas–Andarax Basin. What little vegetation exists here closely follows the river networks which drain southwards into the Andarax. In contrast, the Andarax river valley, which runs from west to east along the northern margin of the Sierra de Gador before turning southwards towards Almeria and the Mediterranean, contains the most noticeable area of dense vegetation. Along this valley cultivation is intense; fruits and vegetables of every kind, especially citrus varieties, have been grown here for hundreds of years.

The sierras are characterized by an even covering of natural upland scrub vegetation, which gives them a pale reddish tinge in the 432 (RGB) image.

**Table 20.8** Multi-temporal datasets used in the case study

Dataset	Scene identifying numbers	Acquisition date(s)
Landsat-5 TM	Subset of path 199/row 034	1984, 1989, 1992
Landsat-7 ETM+	Subset of path 199/row 034	2000
Aerial photography	Ortho-quads (Junta de Andalucia air survey)	2004
SRTM DEM	90 m spatial resolution	2001
ASTER DEM	Subset area of Landsat path 199/row 034, 30 m spatial resolution	2001

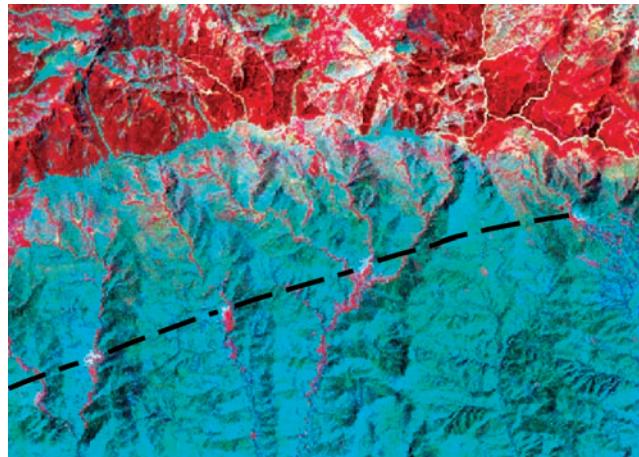


**Figure 20.26** Colour composite of bands 432 DDS (2000) revealing the extent and distribution of vegetation across the Tabernas–Andarax Basin, in red tones. The darker, more contiguous areas of red represent managed upland forest. The remainder of the upland areas are covered by scrub vegetation which appears as a pale reddish tinge

The effect of topography on this natural vegetation distribution can be clearly seen across the watershed of the Sierra de los Filabres in the northern part of the study area. On the south-facing slopes of the sierra (south of the watershed) any kind of vegetation seems extremely sparse, whereas the north-facing slopes are covered with healthy vegetation and in many places this local climatic effect of topography allows the successful growth of managed coniferous forests. The geology and hydrology of the Sierra de Los Filabres is rather different from that of the Sierra de Gador to the south. Here the dominance of relatively impermeable rocks and south-facing slopes means that evapotranspiration is high, whereas porosity and rainfall infiltration are low. These south-facing slopes are largely barren and devoid of healthy vegetation. On closer inspection, the sparsity of vegetation on the southern side can be seen to be punctuated by several tiny flushes of healthy vegetation in valleys, high on the flanks of the mountains. At these locations several villages can be seen, in small white patches in Figure 20.27, on the slopes, about 1000 m above sea level and in a

linear arrangement. Below each village, there is noticeably more healthy vegetation than above. Given the altitude, climate and lack of aquifer here, we conclude that the villages only survive because of a perennial supply of water. We also notice that the villages are aligned and therefore interpret that they are located where the valley floors are intersected by a fault or fracture (black dashed line in Figure 20.27) which brings water to the ground surface at natural springs.

The Sierra de Gador is composed dominantly of dolomite, with some phyllites and schists. The dolomite is fractured and karstified and presents considerable secondary porosity; as mentioned earlier, it forms the major aquifer here. The river flows from west to east and south, along the northern margin of the Sierra de Gador, and into the Mediterranean at Almeria. A very noticeable strip of healthy vegetation can be seen along its length and on the lower reaches of some of its tributaries. Again we must consider what water source is great enough to support such prolific cultivation here. The reason is that, at this margin, water held in the dolomitic basement aquifer is



**Figure 20.27** Landsat 432 DDS subset of the image shown in Figure 20.26 (north-eastern corner). The Sierra de los Filabres villages of Olula de Castro, Castro de Filabres, Velefique and Senes (from south-west to north-east) appear in white on the largely barren south-facing slopes, high up and in a linear arrangement. There is noticeably more healthy vegetation below each village than immediately above, suggesting a structurally controlled water pathway. Notice also the dense natural vegetation (and forestry) on the northern side of the range's watershed. Field of view is ca 12 km

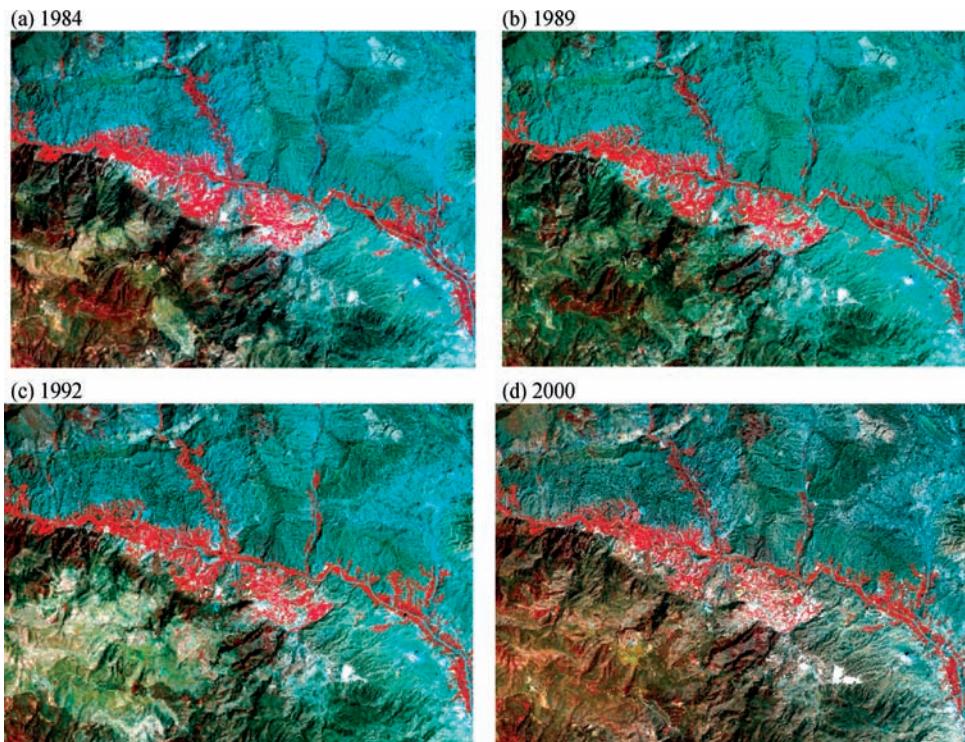
forced to the surface at numerous fault springs and then feed into the Andarax. The simple model, for the regional geology of this area, is of a classic half-graben, with the southern margins of all the basins being faulted and the northern margins being gently sloping. This is of course an oversimplification but it is a useful model that fits quite well here.

We also notice that no such cultivated river valley exists on the gently sloping southern margin of the Sierra de los Filabres. The reasons for this are complex and rely at least partly on the fact that the lithologies are dominantly impermeable so that no large-volume aquifer exists on that side of the basin. Thus the plentiful water here could be considered an accident of geological evolution.

Looking at the standard false colour composites of the Andarax Valley in detail, it can be seen that there has been little change in the extent of cultivation along the valley over the past 20 years or so. The image of 2000 (Figure 20.28d) reveals the appearance of some plastic greenhouses replacing open cultivation but no substantial change in the geographic extent of the cultivated area. The town of Alhama seems also to have expanded in recent years, perhaps because the success of

cultivation has meant increased prosperity, attracting growth and development. Some of the land in and around the town, which was clearly cultivated in 1992, now appears to be urban or devoted to greenhouses.

Closer inspection of the central part of the Tabernas–Andarax Basin, using a calculated NDVI (Figure 20.29), reveals a thin covering of vegetation which appears to be relatively less dense or less photosynthetic, or both. This represents the natural scrub which is characteristic of this and most semi-arid areas. The covering is patchy and its distribution is affected by small-scale topographic and lithological (porosity) variations and, in areas of unconsolidated materials, by potential vulnerability to surface erosion. The highest values (white) appear where there is cultivated vegetation and forested areas, followed by north-facing slopes of the sierras which are populated by healthy natural vegetation (pale grey), with the remaining low-lying basin areas being populated by a very thin, chlorophyll-poor covering of scrub vegetation (darkish grey). The newly completed motorway link to Granada and the main channel of the Andarax river valley show the lowest values (black).



**Figure 20.28** Standard false colour composites (bands 432 DDS) of the immediate Andarax river valley and the lower reaches of the main tributaries, showing the extent and distribution of vegetation between 1984 and 2000; there has been very little change in this time

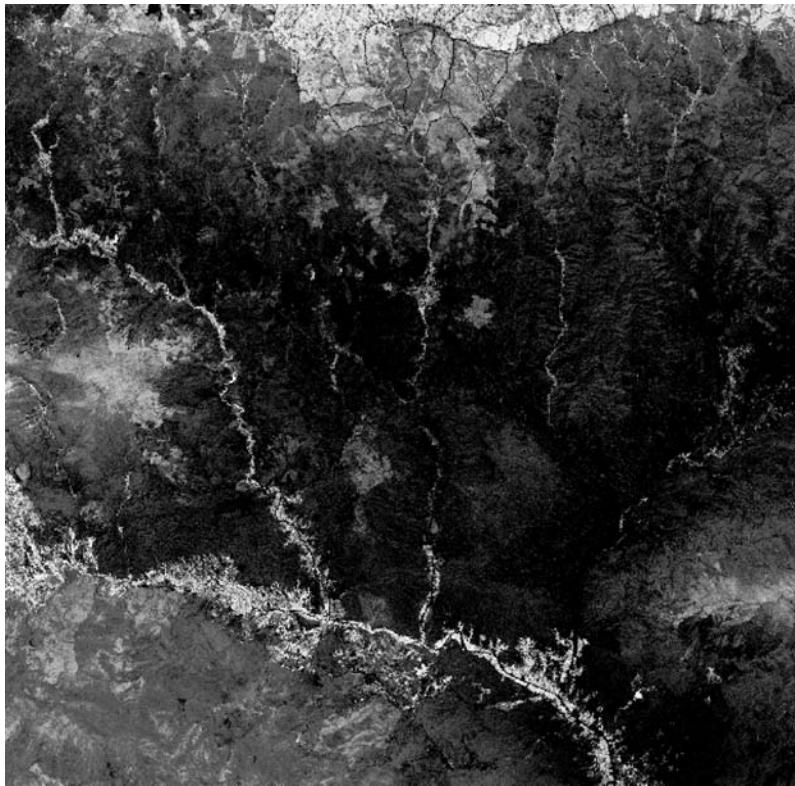
#### 20.4.5 Lithological enhancement and discrimination

Now we draw attention specifically to the geology, although, as we can see from our observations of vegetation above, it is difficult to avoid the geology since it has a controlling influence on many natural phenomena in this area.

If we begin once again with simple colour composites and the general enhancement of rocks and soils, we may start with a simulated true colour image (Figure 20.30) to visualize the main targets of interest as we would see them with the naked eye. We notice that the sierras appear in dark greys, browns and bluish tones, with the sediments and soils of the basins appearing in paler greys, dull cyan, buff brown and greenish tones. We can make some simple but confident divisions of these broad classes from this image alone, as shown in Figure 20.30. Urban areas, buildings and greenhouses appear near white and vegetated areas appear

in very dark grey. These make up the majority of the area. Noticeable against this background are a few areas of reddish brown, the largest of which lies to the west of the town of Gergal. The reddish colour in this image indicates higher reflectance in band 3 and low reflectance in bands 2 and 1, and represents something which looks red to the naked eye.

If we also look at a colour composite of bands 531 DDS (Figure 20.31), which as we know is often the best combination for geological discrimination in semi-arid areas, we are presented with a very useful image revealing many of the significant lithologies very clearly in a variety of vivid colours. Looking first at the Neogene and Quaternary basin sediments, we see that they produce quite a complicated pattern of colours and textures in the centre of the image. The sierras appear as clearly distinguishable on the basis of both tone and texture: the dolomites appear in reddish brown tones while mica schists appear in bluish purple tones. The Tabernas–Andarax Basin is filled with Neogene and Quaternary sediments and

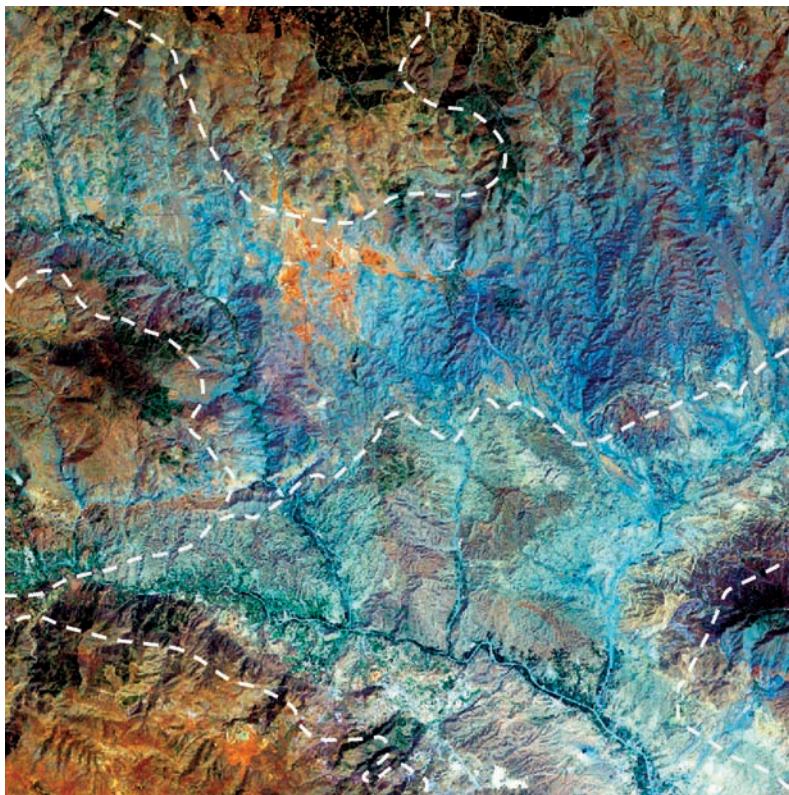


**Figure 20.29** Grey-scale image representing the NDVI, as calculated from the multi-spectral data of the Tabernas-Andarax Basin. The highest values (white) appear where there is cultivated vegetation and forested areas, followed by north-facing slopes of the sierras which are populated by healthy natural vegetation (pale grey), with the remaining low-lying basin areas being populated by a very thin, chlorophyll-poor covering of scrub vegetation (darkish grey). The newly completed motorway link to Granada and the main channel of the Andarax river valley show the lowest values (black)

these appear in a variety of green, grey, blue, pinkish, pale brown and yellow–brown tones. Some division of these lithologies and a hint of folding are interpretable from this area (as shown in Figure 20.31). In this band combination, vegetation appears in very dark reddish tones because it too has relatively low reflectance in bands 3 and 1 and high reflectance in band 5, but less high than the rocks and soils which have their reflectance maxima at these wavelengths. We notice that in this image, the patch of soils to the west of Gergal (appearing red in the previous Figure 20.30) appears in bright olive green tones, indicating high reflectance in band 3 with lower reflectance in band 5 and very low reflectance in band 1.

Looking at this area in more detail, and referring back to our 432 standard false colour composite

(Figure 20.26a), we see that what appears red in the 321 true colour image also appears in greenish tones, again indicating high reflectance in band 3 (green colour gun here) but this time low in both bands 4 and 2. This reflectance pattern is characteristic of iron oxide and hydroxide minerals (such as haematite and goethite), which also have a typically reddish appearance to the naked eye. The area also has a noticeably smooth texture and appears relatively flat compared with the surrounding rocky hillsides, and we may conclude that it represents a pocket of accumulated sediment or soil. Generation of an iron oxide ratio image using bands 3/1 (Figure 20.32), indicates several connected patches with high iron oxide content, relative to the surrounding rocks and soils. These patches of ground suddenly become rather

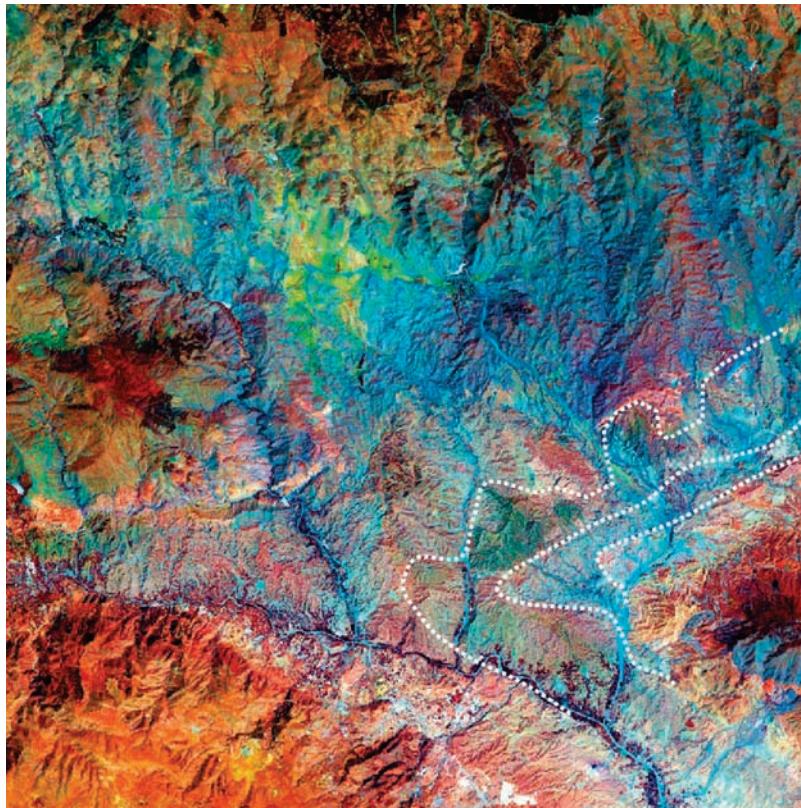


**Figure 20.30** Landsat-7 ETM+ simulated true colour image (321 RGB, DDS) of the Tabernas-Andarax Basin, showing exposed metamorphic basement lithologies in brown (dolomites) and bluish tones (mica schist), with unconsolidated Neogene and Quaternary sediments in cyan grey, buff brown, pale brown and greenish tones. The boundaries between these appear quite clearly and have been indicated by dashed white lines. Vegetation appears dark green and urban areas appear near white, as do areas which have been cleared for cultivation or greenhouse construction

interesting: why do we have a series of isolated areas of reddish soils at this location (rich in iron oxides) and not elsewhere?

Examination of the ASTER 30 m DEM of this area, as shown in Figure 20.32d (as a shaded relief image), reveals two breaks in topography, one to the north of the soil patches and a more subtle one to the south. The soil patches appear aligned in a NW–SE direction and each is elongated in a north–south direction. Careful geological interpretation made from these images and the DEM suggest a series of alluvial fans which drain to the south from the Sierra de los Filabres and are trapped by the topography (see also the photograph in Figure 20.33). They lie along a series of sub-parallel NW–SE-oriented faults; these are related to the main basin-forming

fault systems. The northern part of these faults has uplifted the land to the north, and the southern one has uplifted to the south, producing a small graben between the two. Comparison of the colour composite of bands 531 with the DEM (Figure 20.32b and d respectively) reveals that the basement lithologies in purple and brown tones correspond to the topographic highs. This suggests a rather classical model of uplift-induced erosion and transport of debris from a mountain front and into a fault-controlled topographic trap, producing isolated pockets of sediment. The next question is the reddish colour. The source area of the debris in these pockets lies immediately to the north and consists largely of mica schists, which contain a lot of biotite, and this breaks down very readily and



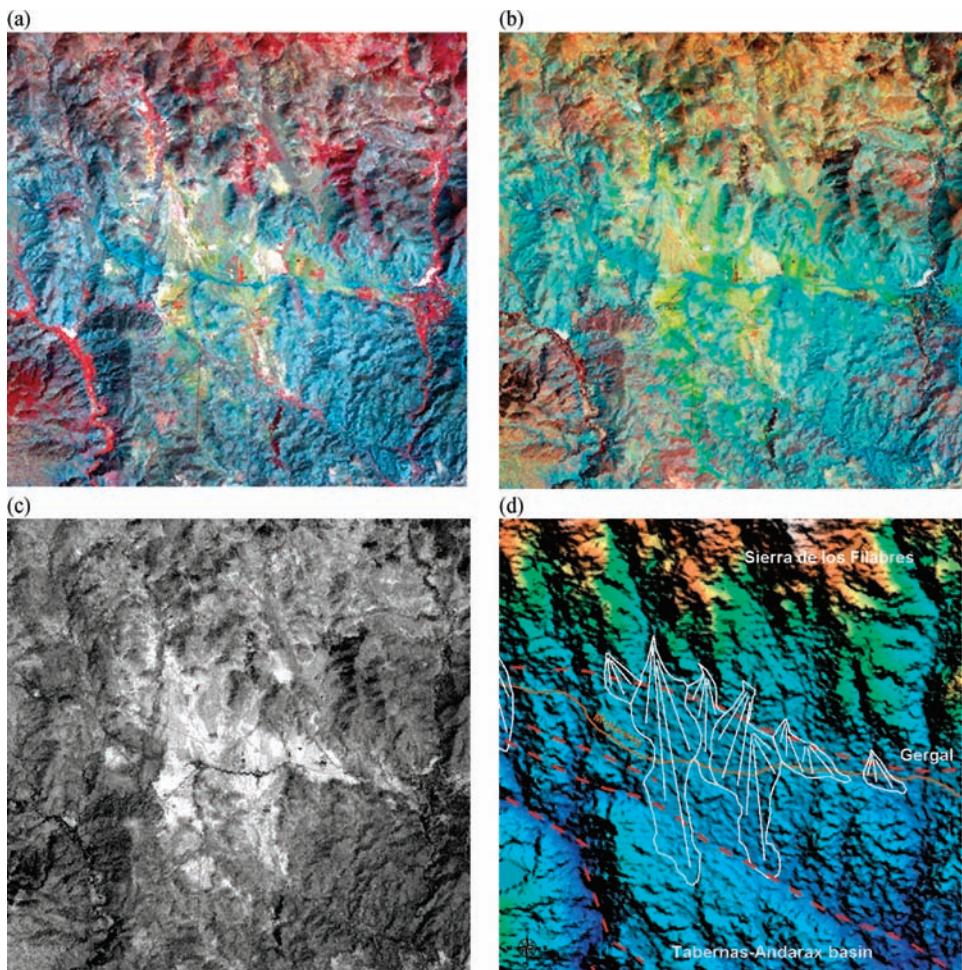
**Figure 20.31** Colour composite of bands 531 DDS for general geological discrimination. In this image, the dolomites and mica schists of the basement massifs appear in pinkish/reddish brown and darkish blue tones respectively. Neogene and Quaternary sedimentary rocks appear as a variety of pinkish, buff brown, greenish, pale bluish grey and yellow–brown tones (interpreted bedding traces have been indicated in dashed white lines). Vegetation appears in very dark red and Quaternary red soils in bright olive green tones

may release its iron to the soils. The unconsolidated nature of the transported debris would then facilitate rapid oxidization of the iron to give the red colour; this may explain the presence of the iron oxides in these soils.

The next question relates to the boundaries between the lithologies of the sierras and basins – can we see these? The answer is yes, but the contacts are not always crisp, linear or well defined. We have already mentioned two of these (and that they are mainly structural): the faulted contact to the south (between the Andarax Valley and the Sierra de Gador); and the gentle dip slope rising northwards to the Sierra de los Filabres. One other forms the northern boundary of the Andarax Valley near Alboloduy and this is described in the next section

#### 20.4.6 Structural enhancement and interpretation

One rather conventional method of enhancing structural features in remotely sensed images comprises spatial filtering, using one or more kernels of varying form and dimension to enhance or suppress features of varying orientation. Choosing which band choice a multi-spectral dataset to use presents a further question. We might prefer to enhance the band or dataset which has the highest spatial resolution. In the case of the Landsat-7 ETM+ dataset, this would be the 15 m panchromatic band. In the case of Landsat-5, however, with no 15 m band we must choose one of the VNIR bands; band 2 or 3 would constitute a good choice since these are less affected by haze than band 1. The reason for spatial



**Figure 20.32** Images of the Gergal area: (a) 432 RGB; (b) 531 RGB; (c) iron oxide ratio of bands 3 and 1; and (d) ASTER DEM shaded-relief image overlain by a simple interpretation of the fan systems (dashed red lines indicate faults, solid white lines indicate alluvial fan systems, the motorway is shown in brown for reference)

filtering in medium-resolution images is that we often cannot see those structural features of interest and so need a little textural help. We use filters to enhance information of different frequencies, high or low. For structural features we may use a high-pass filter to enhance systematic changes in image tone and contrast in the hope of detecting the surface expressions of faults and fractures. At the opposite extreme, filtering of a very high-resolution image, such as the 2004 aerial photography, would seem rather pointless in this sense, since it already gives us unprecedented detail of ground surface features and we can interpret even quite small-scale faults and fractures directly (refer back to the table of

image mapping scales in Section 19.1). In fact a VHR image may overwhelm us with so much spatial detail that we can no longer see the really significant regional-scale structures, so clearly there is a balance to be struck and we need to think carefully about what we are trying to achieve when we (i) choose the data and (ii) decide how to process them.

The structural trends in the study area are complex and the dominant features lie largely on an east–west orientation. These structures comprise the major basin-forming normal faults which were opening and lifting the sierras out of the seas some 15 million years ago. Several other



**Figure 20.33** Red-coloured alluvial fans near Gergal, looking south towards the flanks of the eastern Sierra Nevada (rising to the right or west) and the Sierra de Gador (almost invisible in the far distance, centre left)

older compressional structures (thrusts) exist within the basement complexes of the sierras and these also have an approximately east–west orientation. There are several other structural trends in the basin sediments, produced by Quaternary and post-Quaternary faults on approximately north–south and NW–SE orientations. These are largely normal or oblique-slip faults where the dominant slip vector is vertical.

Highlighting the older basement structures using directional filters is problematic since they are commonly thrusts, which are low-angle structures and produce distinctly nonlinear, rather sinuous surface expressions. They tend to be highlighted more easily by outcrop (lithological) variations and relationships and so are best interpreted visually. This is certainly the case along the south-eastern margin of the eastern Sierra Nevada, near the town of Abolodruoy (Figure 20.34c). Here the surface trace of several thrust faults is highlighted by the presence of relatively highly reflective Neogene sediments (marls containing gypsum). These have been thrust up, as slices between metamorphic rocks (phyllites), from the basin onto the flanks of the Sierra. They form an eye-catching bright east–west-oriented stripe of ground, the uppermost edge of

which is characteristically sinuous, hinting at its low-angled relationship with the hillside. The aerial photographic subset in Figure 20.34c provides great detail, at visible wavelengths, of ground tones and textures but only very subtle tonal differences between the Neogene sediments and the phyllites. In such situations, both spectral detail from Landsat and spatial detail from the digital photograph are vital to the understanding of structural context.

Thus, we must concentrate on the basin-bounding normal faults, and the Quaternary normal faults. Given that there are two dominant trends, east–west and NW–SE, we could use Sobel filters, or variants thereof, to pick out selectively surface topographic features which may indicate structural control. Alternatively, we may choose a Laplacian filter, to highlight any high-frequency textural information and yield a generally sharpened image, or gradient filters to enhance features of a particular direction that we know to exist. The results of such experiments are illustrated in Figures 20.35 and 20.36. The use of a  $5 \times 5$  Laplacian filter yields the image shown in Figure 20.35; this image is effective at highlighting drainage patterns, watersheds and land-use changes. The image is texturally complex and subtle textural changes can be seen across some

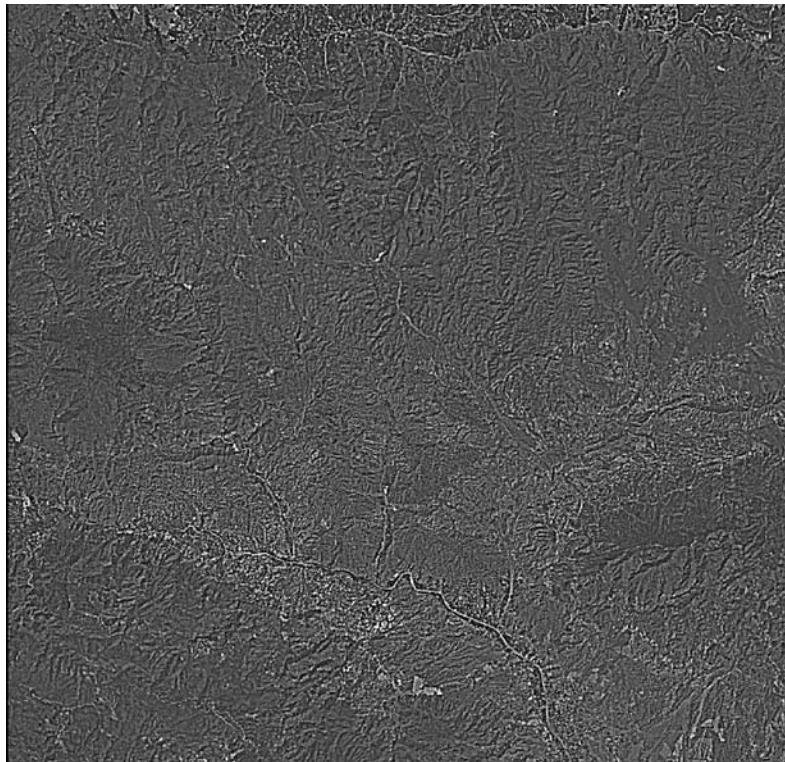


**Figure 20.34** Thrusted contact near the town of Alboloduy: (a) colour composite of bands 531 DDS showing the slice of thrusted basin sediments (marls and gypsum) in bright yellow and orange tones surrounded by mica schists (blues) to the north and other basin sediments (cyan, red and greenish tones) to the south (the black box indicates the coverage of the aerial photograph in (c)); (b) photograph of basement metamorphic rocks and basin sediments thrusted to the north-west (left), looking north-eastwards along the line of the thrusted contact; and (c) 2004 aerial photograph showing VHR detail of the thrust belt near the town of Alboloduy (the red spot shows the location at which the photograph in (b) was taken)

of the larger structures, but otherwise the result proves unhelpful in showing features that we do not already know to exist. Directional Sobel filters yield slightly more promising results as the images in Figure 20.36 show. These images represent the use of a simple  $3 \times 3$  Sobel filter (of the form described in Chapter 4) to highlight features oriented north-south (Figure 20.36a) and east-west (Figure 20.36b). In these images, the main basin-bounding faults can be discerned, though only because we already know they are there. The most

conspicuous features are the bed of the Andarax and its main tributaries, and the watershed of the Sierra de los Filabres.

By far the most effective method of interpreting and so extracting structural features is to use a DEM. Since the features we are looking for are sub-surface phenomena, only some of which may intersect the surface, the main way we can interpret them is by looking at the physical surface to identify systematic topographic expressions which may have structural significance, and hopefully by correlating

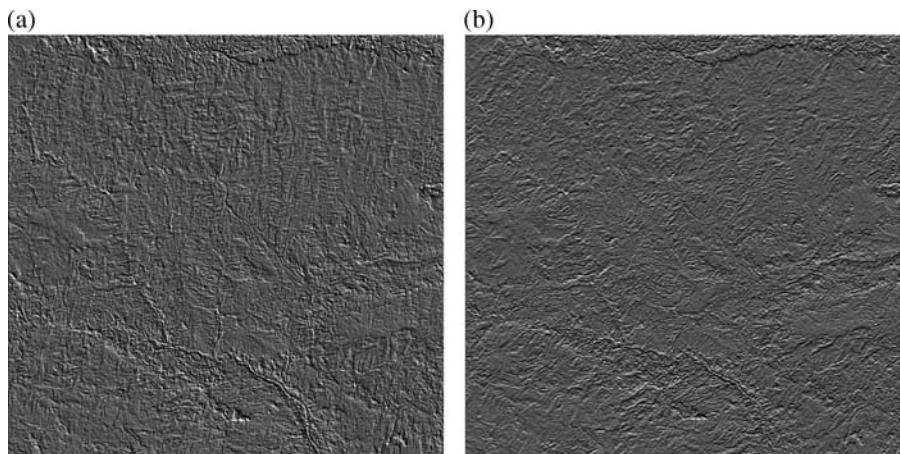


**Figure 20.35** (a) Landsat ETM + band 2 enhanced using a Laplacian filter 3x3. This filtered image is most ineffective for the detection of faults and fractures. Only the most systematic and extensive features are detectable, such as the WSW–ENE trending ridge through the Tabernas Basin which is clearly visible. Otherwise the drainage, vegetation and urban areas near Alhama are highlighted, as is the sharp boundary between the Sierra de Gador and the Andarax Basin sediments

some of them with image spectral variations to make an interpretation. Although the image spectral variations along the boundaries between these terrain units seem subtle and complex, when we focus on the DEM we are able to see several important pieces of information. The Tabernas–Andarax Basin appears clearly in the low-lying areas (darker blues in the DEM shaded-relief image) while the high ground of the sierras is visible in bright oranges. The relatively abrupt change in slope between the Sierra de Gador and the Andarax Valley can be seen in contrast with the gradual northward slope up to the Sierra de los Filabres. We also begin to see the subtle topographic changes of the alluvial fan systems at Gergal (and Tabernas) in addition to their location with respect to the regional structure, as shown in Figure 20.37. These are evidence of relatively recent tectonic activity.

If we then calculate the slope angle (in degrees in this case) from the DEM surface, we can exaggerate these expressions to see them rather more easily (as shown in Figure 20.38). The main basin-bounding faults and many others are revealed by systematic and relatively abrupt changes in gradient. In contrast, the areas occupied by recent alluvial fans are characterized by slope angles of less than 10° and, in the central part of the fan systems, less than 3°.

If we then combine all the fragments of geological knowledge we have gained so far, we can produce a simple regional litho-tectonic interpretation map. This presents the main lithological groups, the major structural elements, recent depositional features (tectonically triggered alluvial fan systems) and the regional aquifers together, thus providing us with a tool to understand better the

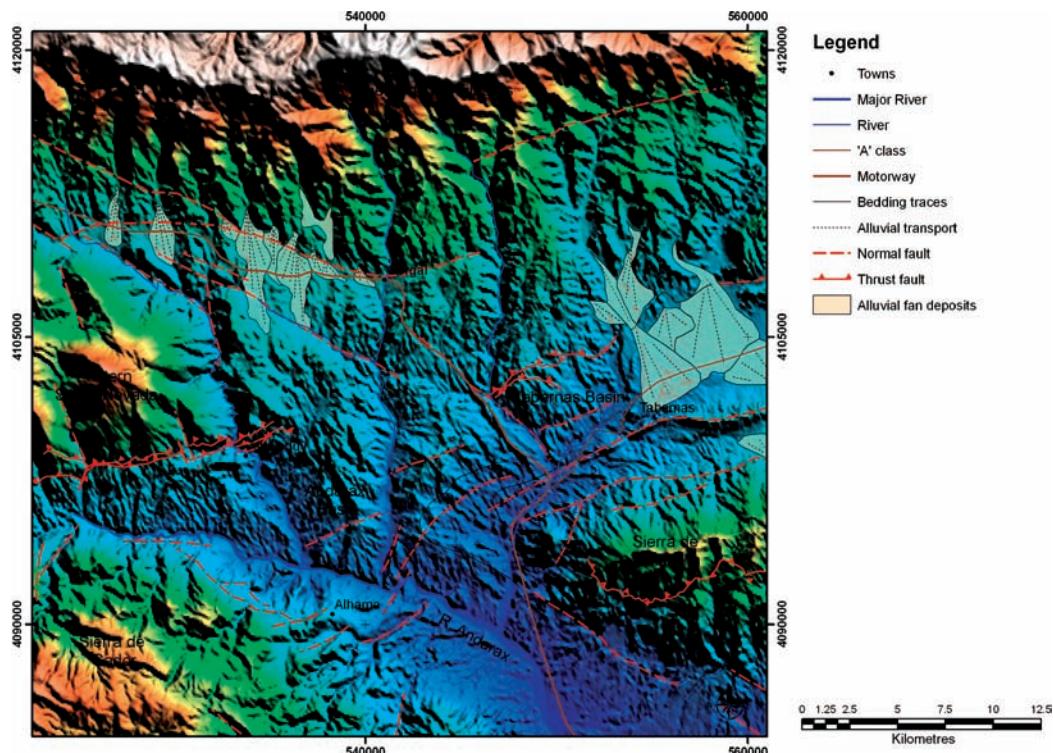


**Figure 20.36** Landsat ETM+ band 2 enhanced using Sobel filters to show linear features on (a) N-S and (b) E-W orientations. Many of the main basin bounding faults are highlighted in these images

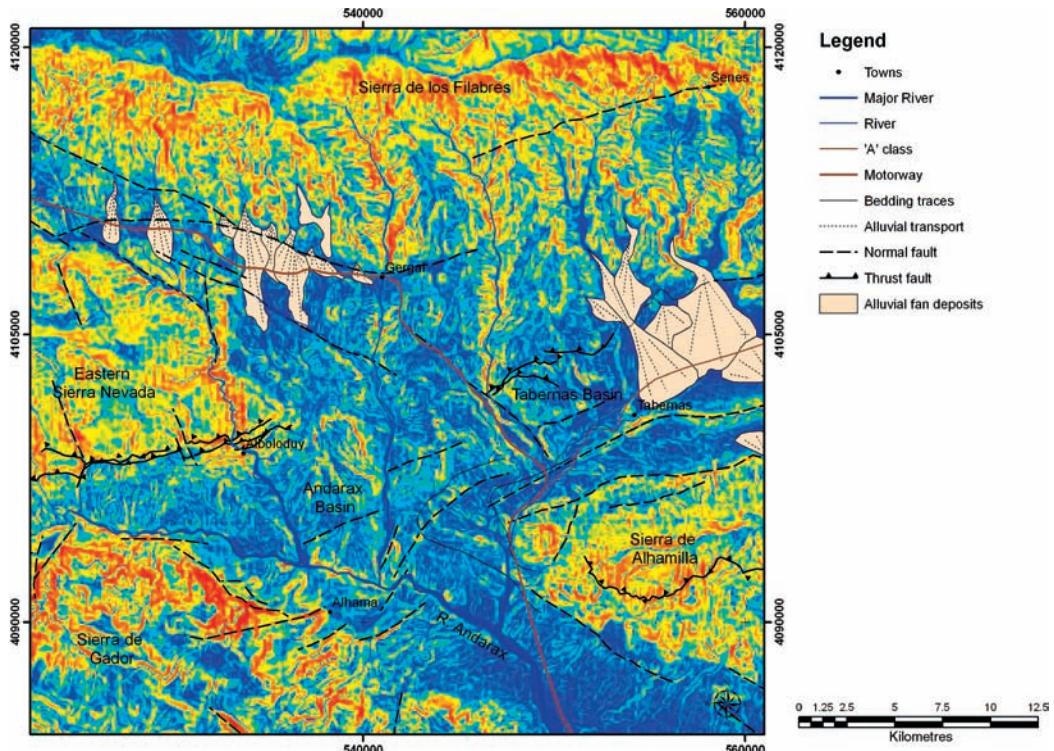
connections between topography, geomorphology, geology, water and agriculture/land cover (land use). The result of this compilation is shown in Figure 20.39.

#### 20.4.7 Summary

This case study is a rather good example of one where we let the images speak for themselves in



**Figure 20.37** SRTM DEM shaded-relief image of the Tabernas-Andarax Basin, with interpreted structures and alluvial fan systems overlain

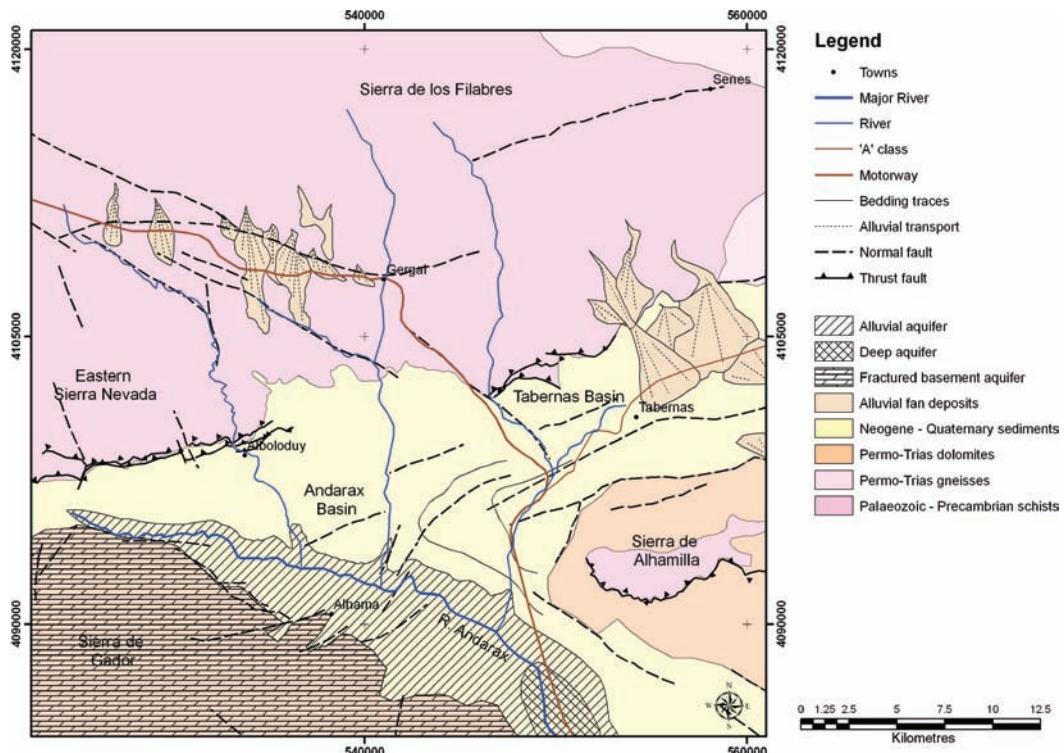


**Figure 20.38** Slope (gradient) as calculated from the SRTM DEM, in degrees, and displayed using a red–blue colour lookup table. Interpreted structures and the alluvial fans interpreted earlier are shown

guiding us towards features of potential interest, rather than approaching the work with a very fixed agenda. Both approaches have their place but when faced with an unknown area and a remit of exploration and understanding, it is often sensible to begin by looking at the most obvious, eye-catching features and then proceeding gradually to the more complex issues. It has also become very clear that water and geology are intricately linked in this area, so that one cannot understand the former without considering and discovering the latter. It is a complex area and so we have attempted to cover only a few of the more interesting ones here.

We find that convolution filtering is not always useful or effective in identifying faults from imagery – you will enhance many surface edges and textural boundaries which bear no connection to sub-surface structures; you may be lucky enough to find some faults. The point to be learned is that filtering should be used when you have a very clear objective and one which is related to purely surface targets. In fact, finding faults in images is more

about using your eyes and applying your geological knowledge than about processing images. Quite often it is a matter of what you do not see that may point to the existence of sub-surface control. For instance, the presence of the three villages, aligned and high up in the mountains without obvious means of supply, could point to a leaky mains pipe but in this environment is more likely to indicate the presence of a spring line. Your understanding of its geological or geomorphological context will then be required to determine whether that spring line is controlled by underlying stratigraphic or structural features. The observation of topographic expression along that line may not be sufficient indication in itself since the expression could still be produced by either form of control. Clearly quite a bit of detective work is required. Only when you have spectral, topographic and stratigraphic agreement can you be reasonably sure that you have identified a fault; however, here again your knowledge and experience will come into play. In any case, you will certainly want to make a field visit to the location to



**Figure 20.39** Summary regional geological interpretation, bringing together lithological and structural information interpreted from the image data and DEM, to produce a regional litho-tectonic map; this can then be used for better interpretation of natural water resources. The aquifer types shown here are those described earlier in the text. Modified after Pulido-Bosch *et al.* (1992)

satisfy yourself and to understand better the geometries of the features on the ground.

The farmers here become victims of their own success since increased demand for their products leads to massive expansion, as in the Níjar area, and increasing pressure on existing resources. This then leads to many questions about sustainable development and the effective management of renewable resources. In this case the cost is not only the reduction of available water but also its quality. Here again, we study a problem retrospectively, leading one to the question the benefit of doing so, given that the real problem has been known about for some time. In doing so, however, we learn many things and begin to understand the necessary approach and methods, and we can then potentially establish a methodology to be applied in other areas and which could be used to give a spatial context to a problem.

## Questions

### Section 20.1

- 20.1 In order to separate gypsum quarries from gypsum natural outcrops, TM3 is added to the last operation of formula (20.2). Can TM1 or TM2 be used instead of TM3 and, if so, why?
- 20.2 Following the logic of formulae (20.1) and (20.2), try to design a ratio-based image processing procedure to enhance gypsum and extract gypsum quarries.

### Section 20.2

- 20.3 The corresponding ASTER and ATM formulae of the compound differencing colour composite in this case study are slightly different.

Try to find the difference between these two groups of formulae and explain why they have been so designed based on the image spectral profiles of argillic alteration, siliceous alteration and gypsum and the spectral bands of the two datasets.

- 20.4 Based on the ATM spectral profiles, explain why PC3 produces high values for gypsum and very low values for argillic alteration zones and the ferroan dolomites.
- 20.5 Explain how gypsum is enhanced in deep blue in Figure 20.15. Comment on the lesson that we can learn from this particular scenario.

### Section 20.3

- 20.6 Very high-resolution imagery provides unprecedented spatial detail of the land and of the greenhouses in this case. Do we really need this detail to carry out the kind of temporal analysis that we have done here?
- 20.7 Why might the increase in plasticulture and the decrease in open vegetation not be reciprocals of one another? What else might be happening?
- 20.8 Without resorting to a ground survey of every individual greenhouse, how else might we improve on this work to estimate greenhouse coverage more accurately?

### Section 20.4

- 20.9 Which other surface parameters could be usefully applied here?
- 20.10 How would you go about estimating how much rainfall enters the groundwater (and the Gador aquifer) here?
- 20.11 What limitations are imposed by the data (Landsat and SRTM DEM) on the interpretation in this case? What data would you advise?

### References

#### Section 20.4

- Gallego, M.C., Garcia, J.A., Vaquero, J.M. and Mateos, V. L. (2006) Changes in frequency and intensity of daily precipitation over the Iberian Peninsula. *Journal of Geophysical Research*, **111**, D24105.
- Pulido-Bosch, A., Sanchez Martos, F., Martinez Vidal, J. L. and Navarrete, F. (1992) Groundwater problems in a semiarid area (low Andarax River, Almeria, Spain). *Environmental Water Geological Sciences*, **20** (3), 195–204.
- Pulido-Bosch, A., Sanchez Martos, F., Navarrete, F. and Martinez Vidal, J.L. (1994) Agricultural practices and groundwater contamination in the Lower Andarax Basin (Almeria, Spain), Water Down Under 94: Groundwater Papers, National conference publication (Institution of Engineers, Australia), no. 94/14, pp. 445–449.



# 21

## Research Case Studies

The chapter is based on the authors' published research papers. The intention of this chapter is not to cover every aspect of remote sensing applications but, instead, using several case studies, to share our experiences with you on the following:

- How to think through and formulate an application research project.
- How to design and develop the most effective image processing techniques and strategy for extracting the required thematic information from images.
- How to establish the most representative and powerful GIS model to serve the objectives of the project.
- How to approach the data analysis and the presentation and critical assessment of results.

### 21.1 Vegetation change in the three parallel rivers region, Yunnan province, China

#### 21.1.1 Introduction

In this case study, multi-temporal Landsat-5 TM and Landsat-7 ETM+ image data were used to assess the change of vegetation coverage. With a simple and effective methodology based on the NDVI, the study aims to identify areas subject to rapid vegetation destruction, as well as to detect any

possible signs of vegetation revival, in the 'Three Parallel Rivers Region' in south-western China (Liu and Meng, 2005).

The area lies within a north-south orogenic belt where the edge of the Eurasian plate is being compressed from the west by the underlying eastward subducting Indian plate. This continental-scale tectonic movement has squeezed and uplifted the terrain dramatically to form the north-south-oriented Hengduan Mountains which lie contrary to the dominant east-west trend of the major mountains further to the north. In this intensely sheared north-south tectonic zone, three great sub-parallel rivers flow in deeply cut valleys separated by high mountains (World Heritage Nomination – IUCN Technical Evaluation, ID No. 1083, 2003). These three rivers from west to east are: Nujiang River (Salween in Burma), Lancang River (Meigong in Vietnam) and Jinsha River (the upper reaches of the Yangtze). The Three Parallel Rivers Region was awarded the prestigious status of 'World Heritage' site by UNESCO in 2003 (World Heritage 27 COM 8 C.4, 2003) for its great diversity of landscape, vegetation, animal species and human culture (natural site datasheet from WCMC). With this new status, the conflict between economic development and environmental protection has intensified. The balance between the two will decide the fate of this rare natural beauty. With abundant water resources and tremendous potential for hydroelectric power, vegetation is a key factor for maintaining a healthy



**Figure 21.1** Location map of the study area in the Three Parallel Rivers Region, Yunnan Province, China

ecological system. Once it is destroyed, severe erosion will occur and the damage will not be restricted to the local environment but will extend further downstream to Burma and Vietnam.

### 21.1.2 The study area and data

The study area is in the north-west corner of Yunnan Province, China, and adjacent to Burma in the west (Figure 21.1). It is within a TM/ETM+ scene of path-row 132-041 extending from  $28^{\circ}6'52''$  to  $26^{\circ}45'48''N$  and  $98^{\circ}23'13''$  to  $99^{\circ}54'53''E$ , covering much of the Three Parallel Rivers Region. The three great rivers are almost parallel to one another in this area; at their closest, they are no more than about 63.4 km apart.

**Table 21.1** Image data of the study area

Image 132-041	Dates (y-m-d)	Temporal separation	Seasonal difference
TM	1994-11-15	6 years, 1 month and 10 days	40 days
ETM+	2000-12-25		

The data were acquired from the Global Land Cover Facility, UMIACS (University of Maryland Institute for Advanced Computer Studies). The TM image used in this study was taken on 15 November 1994 and the ETM+ image on 25 December 2000; the temporal separation between the two is 6 years, 1 month and 10 days, while the seasonal difference is 40 days (Table 21.1). Image co-registration quality is crucial for multi-temporal image comparison. Both images have been ortho-rectified to WGS84 NUTM47 to a high accuracy at source and, as a result, the two images are precisely co-registered without visually observable mismatches even when viewed at pixel level.

### 21.1.3 Methodology

The NDVI is a well-established and robust technique for mapping vegetation based on the diagnostic absorption feature in the red (R) spectrum and very high reflectance in the NIR spectrum (Gausman, 1974; Lillesand and Kiefer, 2000). These two spectral ranges are denoted as bands 3 and 4 in TM and ETM+ image data. One of the advantages of NDVI is that it is normalized to a standard value range from  $-1$  to  $1$  and thus the NDVIs derived from different images are comparable in the same value range. The technique has been widely used for assessment of changes of vegetation status, land-use patterns and ecological parameters (Cihlar, St-Laurent and Dyer, 1991; Lambin and Ehrlich, 1997; Mantovani and Setzer, 1997; Li, Tao and Dawson, 2002; Wang, Price and Rich, 2001).

#### 21.1.3.1 The NDVI difference red, green and intensity composite

The main purpose of the study is the mapping not simply of vegetation but of the changes in

vegetation coverage in the region. To this end, we have composed a simple and effective method to highlight the areas subject to significant change using multi-temporal NDVIs incorporating threshold criteria as described below.

NDVI difference red, green and intensity (NDVI-D-RGI) composite:

$$\begin{aligned}
 \text{Red : } & \text{ If } NDVI1 > C1 \text{ AND } (NDVI1 - NDVI2) \\
 & > T1 \text{ then } NDVI1 \text{ else } NULL \\
 \text{Green : } & \text{ If } NDVI2 > C2 \text{ AND } (NDVI2 - NDVI1) \\
 & > T2 \text{ then } NDVI2 \text{ else } NULL
 \end{aligned} \tag{21.1}$$

Intensity : ETM+ band 4

where  $C$  and  $T$  are the vegetation criterion and vegetation difference threshold. The value range for both parameters is  $[0, 1]$ . The numbers 1 and 2 denote the time sequence of the two images in comparison.

The NDVI-D-RGI composite defined by formula (21.1) produces a vegetation change image. The red layer highlights vegetation destruction (areas covered with healthy vegetation on imaging date 1 but no longer on date 2) in red, while all the unchanged areas, either with or without vegetation on both dates, are output as null. Similarly, the green layer highlights the areas of vegetation revival (no vegetation on date 1 but with vegetation on date 2) in green and leaving all the unchanged areas as null. Overlaying these red and green layers on the ETM+ band 4 intensity layer presents vegetation destruction in red, revival in green and unchanged areas as achromatic imagery background. ETM+ band 4 was chosen as the intensity layer for its high intensity from vegetation and white appearance of snow. Snow appears in black in bands 5 and 7 for its absorption in the SWIR spectral range.

#### 21.1.3.2 Parameter setting for the NDVI-D-RGI

For vegetation comparison, it is vital to acquire the multi-temporal images taken in the same month/season or, better, on the same date/week. Unfortunately, this is not often possible. Small seasonal differences in image acquisition date may produce non-negligible effects preventing a fair comparison for vegetation change assessment. If the image of

$NDVI1$  is taken in a much warmer (or greener) season than that of  $NDVI2$ , a direct comparison between the two images may falsely indicate deterioration of vegetation even if the actual vegetation coverage and conditions are not really changed. Conversely, if the image of  $NDVI1$  is taken in a much colder season than that of  $NDVI2$ , an incorrect conclusion of vegetation revival may be reached. The vegetation criteria,  $C1$  and  $C2$ , and the vegetation difference thresholds,  $T1$  and  $T2$ , in formulae (21.1) allow adjustment to reduce the seasonal bias for vegetation change assessment and control the significance level of the vegetation change to be detected.

The vegetation criteria  $C1$  and  $C2$  decide if an NDVI value is acceptable as vegetation or not, thus eliminating non-vegetation pixels. For TM and ETM+ images, the DNs of vegetation in the near infrared (band 4) should be significantly higher than those in red (band 3), therefore the NDVI of vegetation should always be positive, thus  $C1 > 0$  and  $C2 > 0$  ensure positive values of the NDVI to remove obvious non-vegetation areas. Higher thresholds of vegetation criteria,  $C1$  and  $C2$ , set harsher conditions to reject more pixels from being recognized as vegetation.

The vegetation difference thresholds  $T1$  and  $T2$  set the significance levels of vegetation changes between the two images. For the red layer in (21.1), the vegetation pixels in the  $NDVI1$  image are displayed in red only when their values are greater than their corresponding pixels in the  $NDVI2$  image by a difference of no less than  $T1$ . In this way, pixels showing no significant vegetation change will be eliminated as null. Similarly,  $T2$  will eliminate pixels showing no significant vegetation change in the green layer. Relatively high  $T1$  and  $T2$  thresholds ensure a critical assessment of significant vegetation change (either destruction or revival) while low values of  $T1$  and  $T2$  make an NDVI-D-RGI composite sensitive to changes to both vegetation conditions and coverage. The  $C$  parameters are partially controlled by the corresponding  $T$  parameters. For a given  $T$ , the NDVI difference defined in (21.1) is not sensitive to the variation of  $C$  when  $C < T$ . For instance, for a given  $T1$ , any pixel of  $C1 < NDVI1 < T1$  will be eliminated unless the corresponding  $NDVI2$  has a negative value that makes up the difference of  $T1 - C1$ . The  $C$  parameters only

have strong effects on the NDVI difference when  $C > T$ .

In general, a higher vegetation criterion and a higher vegetation difference threshold should be set for the NDVI image taken in a warmer (or greener) season so as to compensate for the vigorous effect of vegetation. The value of  $C_1$  or  $C_2$  should be set proportional to the seasonal greenness, i.e. the greener the vegetation on the imaging date, the higher its vegetation criterion should be, but this simple principle is not applicable in areas of high relief. The NDVI cannot effectively suppress topography and may yield much lower values for vegetation in dark shadows than on illuminated slopes. A high vegetation criterion ( $C_1$  or  $C_2$ ) removes too many vegetation pixels in areas of dark shadows but a low criterion makes this parameter nearly redundant if  $C < T$ . The vegetation criterion is therefore effective only for seasonal compensation in low relief and flat areas.

More effective compensation can be achieved by setting different values for  $T_1$  and  $T_2$  depending on the seasonal greenness difference between the two images in formula (21.1). A higher vegetation difference threshold should be set to the NDVI image of a warmer (greener) season. For instance, if  $NDVI_1$  is taken in a greener season than  $NDVI_2$  in formula (21.1), then we should set  $T_1 > T_2$ . The difference between  $T_1$  and  $T_2$  decides the strength of the compensation to seasonal greenness bias.

The specific parameter setting and its effects can be adjusted and judged empirically, with reference to NDVI image statistics. Comparisons between the standard false colour composites of the two dates, in conjunction with the NDVI-D-GRI composite, can help to ensure effective parameter settings and accurate detection of evident vegetation changes.

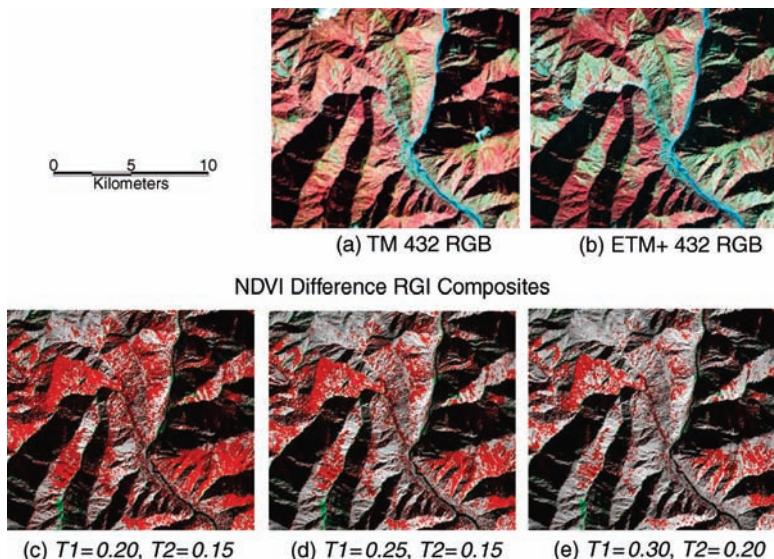
## 21.1.4 Data processing

As shown in Table 21.1, the temporal separation between image 1 (TM) and image 2 (ETM+) is 6 years, 1 month and 10 days while the seasonal difference between them is 40 days. The imaging date of the TM image is in a warmer month, 15 November, than that of the ETM+, 25 December. Without any actual vegetation change, the TM image would appear ‘greener’ or with a higher average NDVI than the ETM+ image. To ensure that the weak signal of vegetation in deep shadows is not eliminated, the vegetation criteria for both images were set to a low positive value  $C_1 = C_2 = 0.1$ . As shown in Table 21.2, the mean and median of the TM NDVI image are significantly higher than those of the ETM+ NDVI image. This is likely caused by both the seasonal bias and the significant reduction of vegetation coverage over 6 years. To compensate for the seasonal effects, the general setting for the vegetation difference thresholds is  $T_1 > T_2$ . Three different sets of  $T_1$  and  $T_2$  with increasing thresholds for vegetation change were applied for comparison (Table 21.2). All these settings are slightly favourable to vegetation revival (green pixels) in the resulting NDVI-D-RGI composite to avoid exaggeration of vegetation destruction.

Linking NDVI-D-RGI composites with standard colour composites of TM and ETM+, we can observe how different settings affect the detection of vegetation change, as illustrated in Figure 21.2. The number of pixels identified as vegetation destruction (red pixels) decreases considerably with increasing  $T_1$  while the number of pixels representing vegetation revival (green pixels) decreases with increasing  $T_2$ . The first set of parameters ( $T_1 = 0.20$ ,  $T_2 = 0.15$ ) with  $T_1 - T_2 = 0.05$  is equivalent to boosting the mean of  $NDVI_2$  by 0.05 and the

**Table 21.2** TM and ETM+ NDVI statistics and settings for parameters C and T

Colour	Images: NDVI	NDVI statistics			$C_1$	$C_2$	$T_1$	$T_2$
		Mean	Median	Std dev.				
Red	TM: $NDVI_1$	0.299	0.327	0.225	0.1	0.20	0.25	0.30
Green	ETM+: $NDVI_2$	0.204	0.205	0.195	0.1	0.15	0.15	0.20



**Figure 21.2** The effects of  $T$  (vegetation difference threshold) parameter setting on NDVI-D-RGI composites: (a) the 1994 TM 432 RGB image; (b) the 2000 ETM + 432 RGB image; (c) the NDVI-D-RGI composite derived from  $T_1 = 0.20$ ,  $T_2 = 0.15$ ; (d) the NDVI-D-RGI composite derived from  $T_1 = 0.25$ ,  $T_2 = 0.15$ ; and (e) the NDVI-D-RGI composite derived from  $T_1 = 0.30$ ,  $T_2 = 0.20$

resulting image in Figure 21.2c can be interpreted as change in both vegetation condition as well as coverage. The second set of parameters ( $T_1 = 0.25$ ,  $T_2 = 0.15$ ) with  $T_1 - T_2 = 0.1$  boosts the mean of  $NDVI_2$  to a level slightly higher than that of  $NDVI_1$ . With the increased  $T_1$  and  $T_1 - T_2$ , the resulting image in Figure 21.2d more critically targets the severe vegetation destruction which caused the decrease in vegetation coverage (see Figure 21.2a and b). The image derived from the third set of parameters ( $T_1 = 0.3$ ,  $T_2 = 0.2$ ) in Figure 21.2e may well represent too harsh an assessment; many pixels showing obvious changes in vegetation coverage in Figure 21.2d were eliminated by the high vegetation difference thresholds  $T_1$  and  $T_2$ .

Apart from the dominant image features relating to changing vegetation coverage, there are several sources of error which produce odd features in the NDVI-D-RGI composites. The snow coverage in the two images varies according to the season and the weather conditions. A vegetated area with snow cover in the TM image but without snow cover in the ETM + image will appear in green in the NDVI-D-RGI composite, meaning an incorrect indication of vegetation revival, because the snow is recognized

as indicating no vegetation in the TM image. The opposite results in a red patch in the NDVI-D-RGI composite, and indicates a similar false alarm for vegetation destruction. The vegetation coverage change detected along the edge of the permafrost zone in the high mountains must therefore be verified carefully. Clouds introduce the same type of errors under the same logic.

Both snow and clouds have their strongest reflectance in the blue spectrum, recorded in TM/ETM+ band 1, and thus can be effectively eliminated using a blue band threshold in NDVI generation as below:

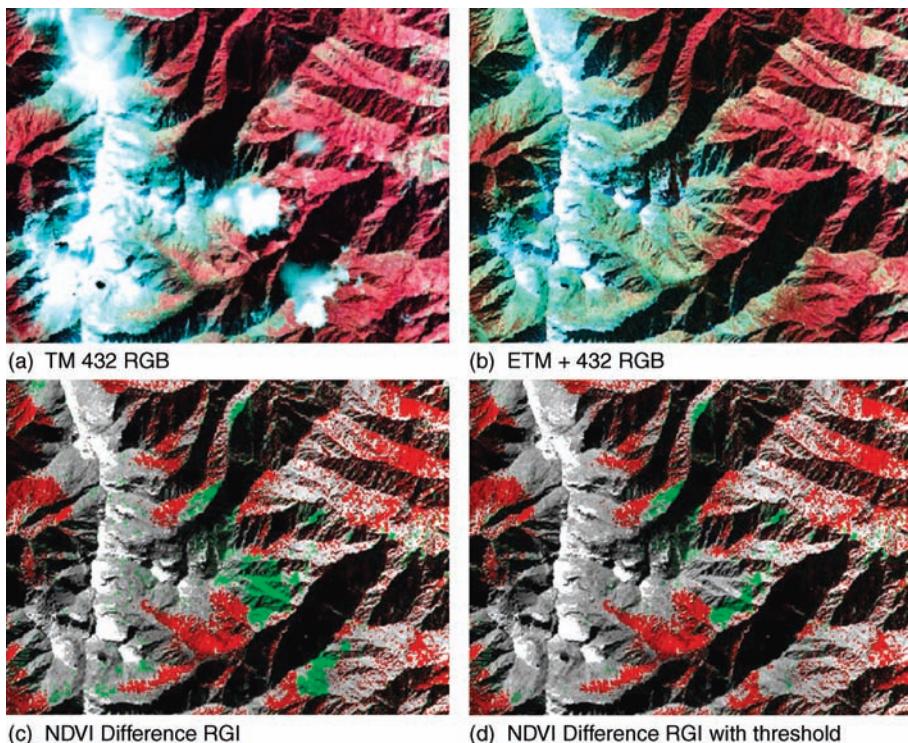
$NDVI_1$  (TM) :

```
If band 1 < 70 then (band 4 – band 3) /  
(band 4 + band 3) else NULL
```

$NDVI_2$  (ETM+) :

```
If band 1 < 80 then (band 4 – band 3) /  
(band 4 + band 3) else NULL.
```

The thresholds for the two images are set slightly different as the average DN level of ETM+ band 1 is slightly higher than that of TM band 1.



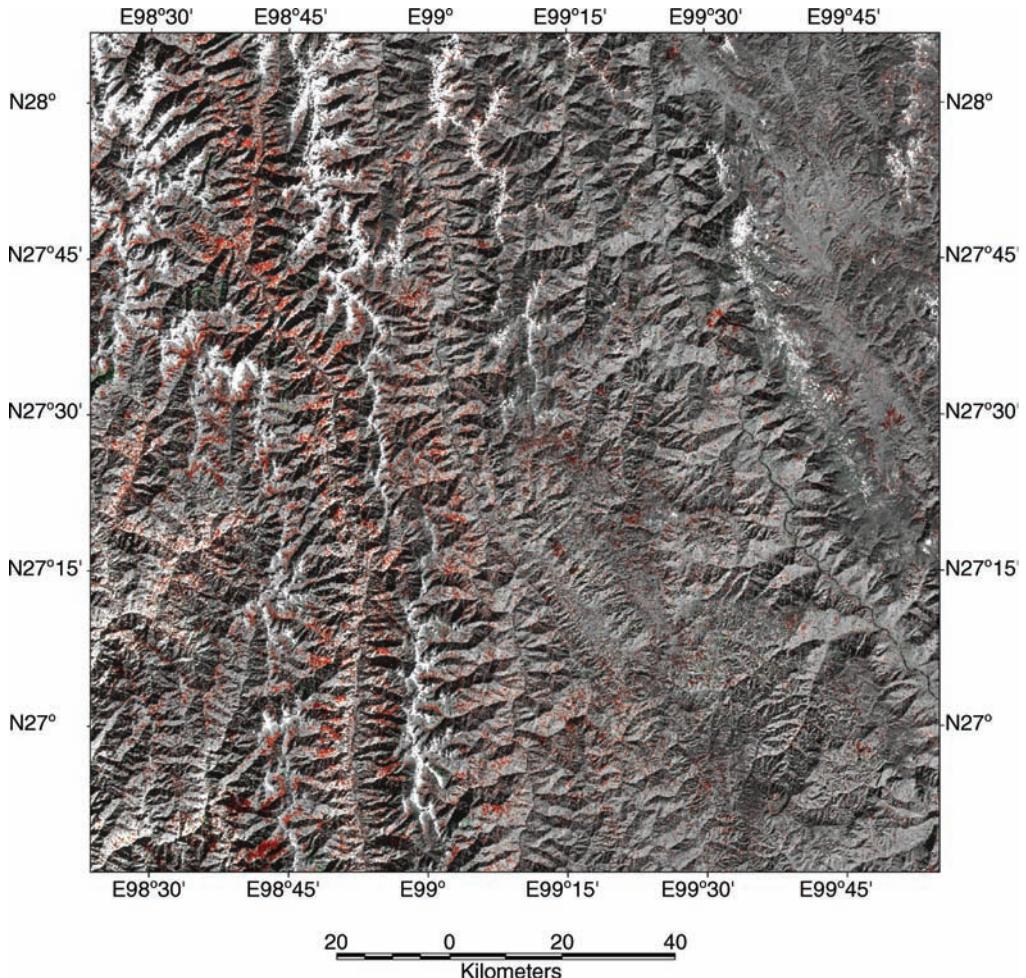
**Figure 21.3** The effects of cloud, cloud shadow and snow: (a) the 1994 TM 432 RGB image where the mountain was covered by both snow and clouds; (b) the 2000 ETM+ 432 RGB image where the mountain was covered by less snow than 1994 and without cloud; (c) the large green patches in the NDVI-D-RGI composite are clouds and cloud shadows in the TM image while the scattered green pixels along the ridge of the snow mountain are caused by snow retreat in the ETM+ image; and (d) after applying the blue-band thresholding, snow retreat and clouds are effectively removed, while there are still some residual green patches produced by cloud shadows that cannot be easily eliminated

As shown in Figure 21.3, the high mountain was covered with both snow and clouds in the TM image (Figure 21.3a) but covered with snow only in the ETM+ image (Figure 21.3b). In the NDVI-D-RGI composite without the blue band thresholding (Figure 21.3c), clouds and some scattered pixels of snow are wrongly recognized as vegetation revival in green. The green patches and scattered green pixels corresponding to clouds and snow are effectively removed in the image in Figure 21.3d by the blue band thresholding, but there are still some residual green patches which are caused by the cloud shadows in the TM image and these cannot be easily removed.

The NDVI-D-RGI composites used for the interpretation of regional vegetation changes in the following section are all shown with the blue band thresholding.

### 21.1.5 Interpretation of regional vegetation changes

Visual observation indicates that the NDVI-D-RGI composite with the second set of parameters (Figure 21.4) provides a well-balanced estimate of the vegetation change. In this image, areas subject to significant vegetation destruction have been effectively identified and there is no obvious evidence for the exaggeration of subdued vegetation features (possible seasonal effects) in the ETM+ image as destruction of vegetation. On the other hand, the image is ‘kind’, even to subtle vegetation revival, and this is ensured by the large difference between  $T_1$  and  $T_2$ . This image is used as the principal image for interpretation. The statistics of vegetation changes derived from the NDVI-D-RGI composites, of the three different parameter settings, are



**Figure 21.4** The NDVI-D-RGI composite derived from 1994 TM and 2000 ETM+ images. The image shows the change in vegetation coverage over the six years between the two imaging dates; red indicates vegetation destruction during the period, green indicates areas of vegetation revival, while the grey-scale background presents the areas which are unchanged

summarized in Table 21.3. Again, the statistics derived from the image of the second parameter setting are used as the basis for discussion while the statistics of the other two images serve as lower and upper limits.

The TM/ETM+ NDVI-D-RGI composite in Figure 21.4 illustrates changes to vegetation coverage during the interval of 6 years. In general, the region is largely covered by natural vegetation (forests, bushes and grass), particularly in mountainous areas. The limited areas devoted to agriculture usually occur in relative flat areas or wide valley bottoms but are not cultivated in winter.

Therefore the vegetation change detected using the winter TM and ETM+ images in this study is little affected by the change of cultivation in these crop fields.

In reference to the whole study area, the widely spread red patches in the NDVI-D-RGI composite indicate that vegetation coverage has decreased rapidly. As shown in Table 21.3, in the 2 255 227 hectare area, 209 999 hectares of vegetated land in 1994 became barren in 2000; the reduction is 9.3%, according to the second set of parameters. Visual interpretation indicates that most of the noticeable green patches are caused by cloud shadows in the

**Table 21.3** Statistics of vegetation changes in the Three Parallel Rivers Region. The numbers in bold are used in the text

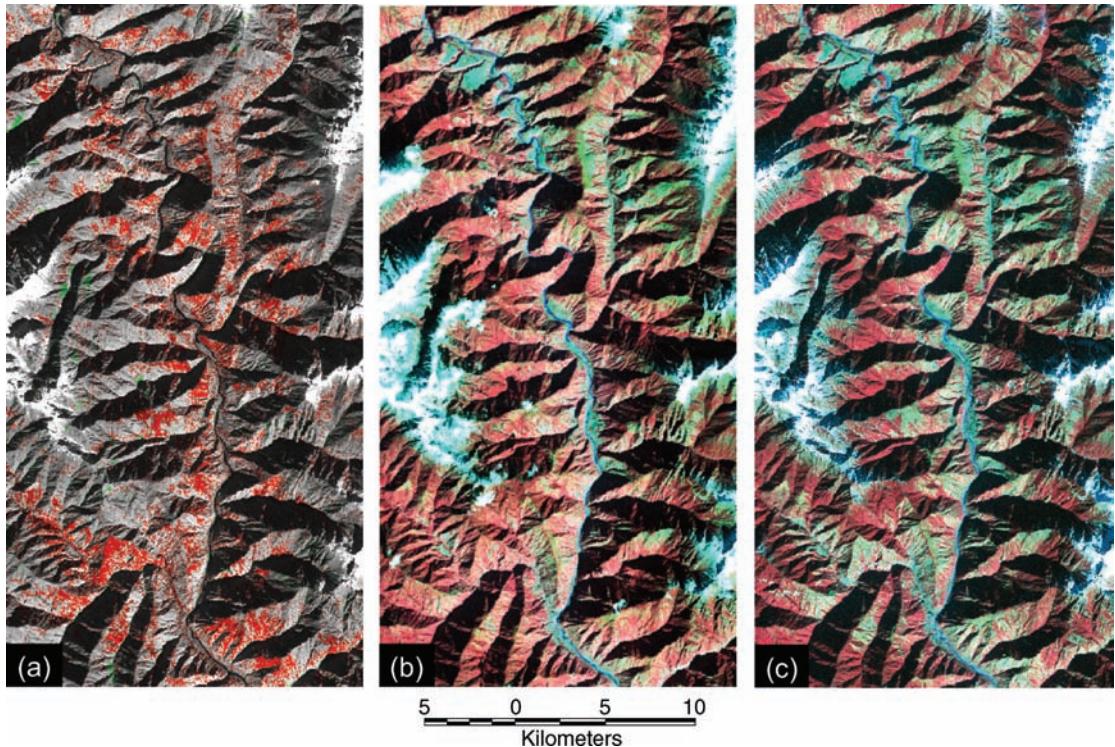
Area name pixels hectares	Parameter		Destruction			Revival			Net (%)
	T1	T2	Pixels	Hectares	%	Pixels	Hectares	%	
Whole study area 27 765 180	0.2 <b>0.25</b>	0.15 <b>0.15</b>	4 442 614 <b>2 585 403</b>	360 851 <b>209 999</b>	16.0 <b>9.3</b>	477 969 <b>477 969</b>	38 823 <b>38 823</b>	1.7 <b>1.7</b>	14.3 <b>7.6</b>
2 255 227	0.3	0.2	1 448 928	117 689	5.2	288 444	23 429	1.0	4.2
Nujiang river region 5 109 595	0.2 <b>0.25</b>	0.15 <b>0.15</b>	1 124 343 <b>727 621</b>	91 317 <b>59 101</b>	22.0 <b>14.2</b>	98 760 <b>98 760</b>	8 022 <b>8 022</b>	1.9 <b>1.9</b>	20.1 <b>12.3</b>
415 029	0.3	0.2	443 325	36 009	8.7	67 492	5 482	1.3	7.4
Lancang river region 5 837 570	0.2 <b>0.25</b>	0.15 <b>0.15</b>	1 092 695 <b>626 844</b>	88 754 <b>50 915</b>	18.7 <b>10.7</b>	85 836 <b>85 836</b>	6 972 <b>6 972</b>	1.5 <b>1.5</b>	17.2 <b>9.2</b>
474 156	0.3	0.2	334 762	27 191	5.7	48 762	3 961	0.8	4.9
Jinsha river region 8 663 020	0.2 <b>0.25</b>	0.15 <b>0.15</b>	1 069 892 <b>560 907</b>	86 902 <b>45 560</b>	12.4 <b>6.5</b>	153 304 <b>153 304</b>	12 452 <b>12 452</b>	1.8 <b>1.8</b>	10.6 <b>4.7</b>
703 654	0.3	0.2	288 746	23 453	3.3	84 170	6 837	1.0	2.3

TM image. Even if we accept these green patches as vegetation revival, the net vegetation destruction would still be 7.6%.

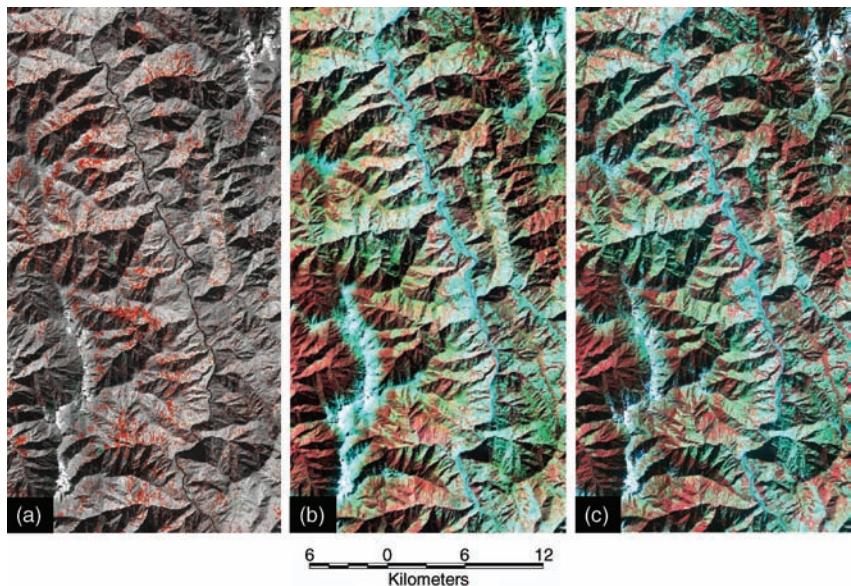
The most stunning features in Figure 21.4 are the concentrated red patches along the Nujiang River indicating an alarming rate of vegetation destruction, particularly in the northern section of the river, as illustrated in Figure 21.5a. The standard false colour composites of TM and ETM+ images (Figure 21.5b and c) show that the Nujiang River catchment was much better covered by vegetation than the catchments of the other two rivers, but better vegetation cover means greater potential for destruction. The NDVI-D-RGI composite (Figure 21.5a) highlights the areas where healthy vegetation existed in 1994 (Figure 21.5b) but no longer in 2000 (Figure 21.5c). The vegetation coverage reduction in the 405 129 hectare area along the Nujiang River is about 59 101 hectares, amounting to 14.2% based on the second set of parameters (Table 21.3). The area also shows a slightly higher revival rate (1.9%) than the regional average (1.7%) but this information is not reliable. Scattered clouds in the 1994 TM image appear mainly in this area which is nearly completely cloud free in the 2000 ETM+ image. As a result, those limited recognizable green patches are nearly all produced by cloud shadows in the TM image (see Figure 21.3). There is no clear evidence of

vegetation revival in the catchments. The severe destruction of vegetation is mainly along the river forming a belt. This may explain the dramatic recent increase in flood and mudflow hazards in the areas further downstream.

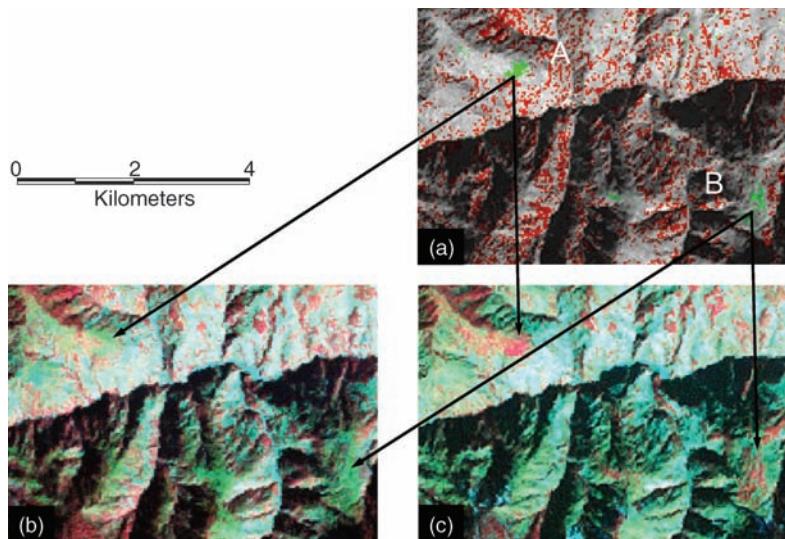
Parallel to and east of the Nujiang River is the Lancang River. The destruction of vegetation in the Lancang River catchments was the worst among the three rivers and its status was already poor in 1994 when the TM image was taken, leaving a lower potential for further deterioration. Consequently, the decrease in vegetation coverage in the Lancang River catchments appears not as significant as that in the Nujiang River catchments but the degradation of vegetation was still severe, particularly on mountain slopes along the west bank of the river, shown as scattered red patches in the NDVI-D-RGI composite (Figure 21.6). The vegetation coverage reduction rate calculated from the second set of parameters for the 474 156 hectare area along the Lancang River is about 50 915 hectares (10.7%), as shown in Table 21.3. The scattered green spots around snowy mountain peaks in this area are, in general, caused by snow cover which has retreated slightly in the 2000 ETM+ image, but a few green patches are noticeable and these indicate new plantations (Figure 21.7). Accounting for all the green pixels as vegetation revival, the net vegetation destruction is 9.2%.



**Figure 21.5** North section of Nujiang River in the study area. The red patches in the NDVI-D-RGI composite (a) indicate devastating destruction of vegetation, which can be clearly seen by comparing the standard false colour composite (b) 1994 TM 432 RGB image with (c) the 2000 ETM+ 432 RGB image. The green patches in (a) are caused by cloud shadow



**Figure 21.6** Middle section of Lancang River in the study area. The NDVI-D-RGI composite (a) illustrates severe destruction of vegetation along the west side of the river where the already poor coverage of vegetation has further deteriorated, as shown in (b) the 1994 TM 432 RGB image and (c) the 2000 ETM+ 432 RGB image in comparison



**Figure 21.7** A new plantation is detected in the NDVI-D-RGI composite (a). The green patches denoted by A and B in (a) were barren in 1994 as shown in the TM 432 RGB image (b) but then covered by healthy vegetation in 2000 shown in red in the ETM+ 432 RGB image (c)

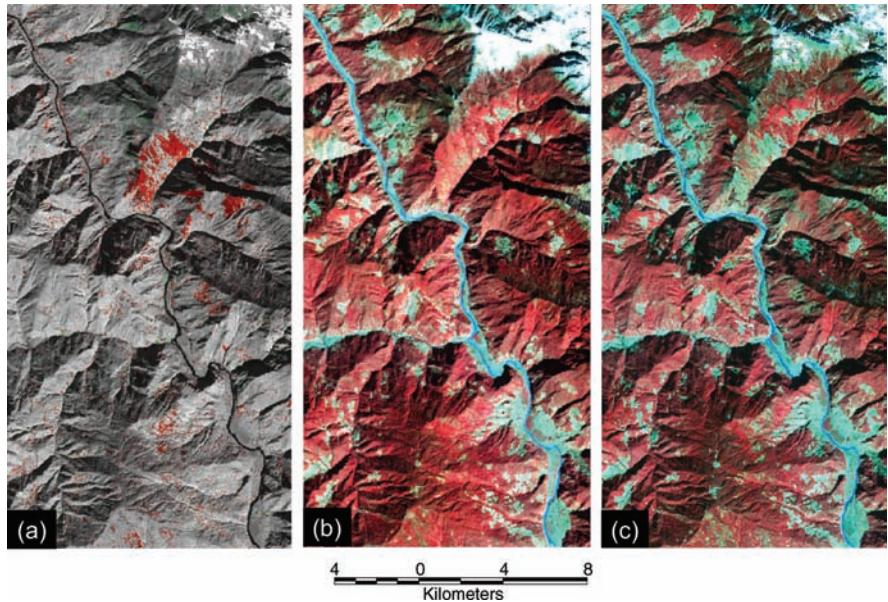
The Jinsha River lies further east of the Lancang River. The Jinsha River catchments in the scene show relatively good vegetation coverage and the least change during 6 years as compared with the other two rivers (Figure 21.4). The general trend of vegetation in the Jinsha River catchments was still in the direction of degradation, as indicated by widely spread red spots in the NDVI-D-RGI composite, as well as a few isolated red patches of obvious vegetation destruction (Figure 21.8). Most scattered green spots in this part of the image indicate a weak revival of vegetation on some slopes (Figure 21.8). Taking all the green pixels as vegetation revival, the net reduction of vegetation coverage in the 703 654 hectares of the Jinsha River catchments in the study area is 4.7%, according to the second set of parameters, which is significantly lower than the areas along the other two rivers (Table 21.3).

### 21.1.6 Summary

The data derived from this study only represent the changes of vegetation coverage between 1994 and 2000. During a field investigation in 2004 and 2006, we noticed that the local government and local

population have made a great effort in planting trees and in protecting the natural vegetation, but their efforts may be cancelled out by the massive developments in road building, and other engineering work, in an attempt to fulfil the demands of rapidly growing tourism since the Three Parallel Rivers Region was granted status as a World Heritage site. The picture in Figure 21.9 was taken in Nujiang Valley in 2006; the massive destruction of vegetation and soil loss are obvious and the damage caused by road cutting was devastating.

As the assessment is sensitive to subjective choices of parameters  $C$  and  $T$ , the accuracy of the statistics of vegetation destruction derived from this study is subject to detailed verification, but the evidence of severe destruction of vegetation and rapid reduction of vegetation coverage shown by this study is unequivocal. The three sets of parameters are all favourable towards vegetation revival rather than destruction. The vegetation destruction in shadowed areas can only be significantly underestimated because of weak signals. We therefore believe that the evidence of severe vegetation destruction based on the second set of parameters is reasonably close to the true situation and is more likely an underestimation rather than an exaggeration.



**Figure 21.8** In the catchments of the Jinsha River, a few obvious red patches in a largely grey background in the NDVI-D-RGI composite (a) reveal limited areas of severe vegetation destruction in an otherwise well-preserved region. In the obvious red patches highlighted by (a), healthy vegetation shown in (b) the 1994 TM 432 RGB image has been completely stripped, as confirmed by (c) the 2004 ETM+ 432 RGB image. The scattered subtle green spots on the slope north of the obvious red patches in (a) imply weak revival of vegetation; the phenomena are reflected by increased redness at the corresponding locations in (c) in comparison with (b)



**Figure 21.9** Field photo taken in April 2006 on the road along the Nujiang Valley. The picture shows the destruction of vegetation coverage by road cutting and excessive cultivation

On the other hand, we realize that vegetation destruction is much easier to detect than revival. Most features of severe vegetation destruction are the result of human activities (such as logging, burning and engineering), which can be produced in a short period and with clear boundaries. In contrast, vegetation revival is a long and gradual process, and its features are subtle, scattered and more sensitive to seasonal effects. With these factors in mind, vegetation revival could be underestimated.

## 21.2 Landslide hazard assessment in the three gorges area of the Yangtze river using ASTER imagery: Wushan–Badong–Zogui

This case study presents a regional assessment of landslide hazard in the Three Gorges area, China, based on Terra-1 satellite ASTER image data, including a stereo-image-derived DEM and multi-spectral reflective and thermal imagery, in combination with field investigations.

A simple, multi-variable elimination and characterization model, employing geometric mean and Boolean decision rules, has been applied to a multi-criterion image dataset to categorize the area into a series of potential landslide hazard levels, which are presented in map form (Liu *et al.*, 2004; Fourniadis, Liu and Mason, 2007a, 2007b).

### 21.2.1 Introduction

The Three Gorges Dam and reservoir project has gained international attention, not only for its great potential for hydroelectric power generation and flood control, but also for its potentially harmful effects on the environment and socio-economy. The most significant and widespread natural hazard in the region is slope instability. There are more than 2500 known localities of slope instability there. With a significant increase and periodic fluctuation of the pool level in the reservoir, the stability of the huge shore area is a grave and unavoidable problem. Slope instabilities already threaten several new towns and the raising of the reservoir level has

the potential to reactivate old instabilities as well as trigger new ones. Here we present an independent study of terrain stability and landslide hazard assessment in the Three Gorges area of the Yangtze, using Terra-1 satellite ASTER imagery and a DEM, in combination with field observations.

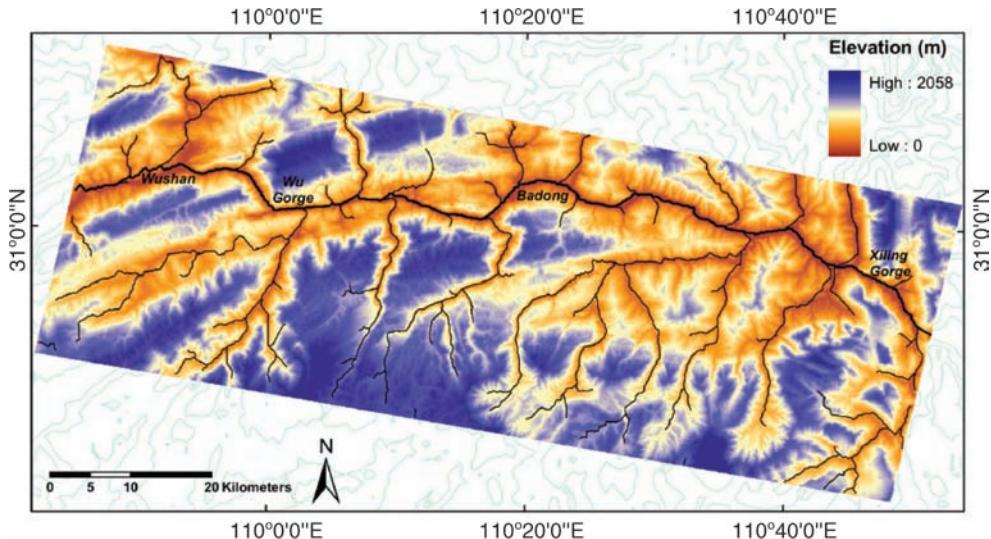
Landslide and slope instability can only be regarded as hazards when they have a negative impact on human life and the environment, otherwise these are part of the erosion process on the Earth's surface to flatten mountains and reform the landscape. In the Three Gorges area, populated regions are often confined to low-lying areas on outcrops of less resistant materials where soils are more mature, which means that many villages and farms are sited on unstable terrain. For this reason, our study is focused on the populated regions along the Yangtze in the Three Gorges area. We also consider the interaction and balance between natural processes and human activities, the disturbance of which is often a trigger for landslides in this area.

We understand that the current development of new towns and new settlements along the Yangtze is constrained by many natural, economic and cultural factors. We intend therefore to produce spatial information about relative levels of hazard, on the basis of integrated digital datasets, which might assist planners and decision makers in ensuring that appropriate engineering measures are taken.

### 21.2.2 The study area

#### 21.2.2.1 Geography

The Three Gorges have been formed by severe incision of massive limestone mountains, of lower Palaeozoic age and Mesozoic age ( $J_1$ , Jialinjiang Group), along narrow fault zones, in response to Quaternary uplift (Chen *et al.*, 1995; Zhao, 1996; Li, Xie and Kuang, 2001; Huang, Xie and Kuang, 2001). Between the gorges, the rocks are much less resistant, consisting mainly of thinly interbedded sandstones, shells and limestone, and the valley is wider with less steep slopes than in the gorges. River bank erosion, terrain dissection and slope failures tend to be concentrated in these 'inter-gorge' areas. Our study area includes the Wushan, Badong and Zigui Counties, between the Wu and Xiling Gorges (Figure 21.10). The annual average



**Figure 21.10** Geographical setting of the study area. The Yangtze River traverses the area, flowing over the county boundary from Wushan (west) to Badong and Zigui (east)

precipitation, in this part of China, in 2002 was 100–150 mm per month and the spring–summer (March–August) average can be as high as 200–300 mm per month (<http://www.dwd.de/research/gpcc>).

#### 21.2.2.2 Geology

The basement of the study area is a crystalline, pre-Sinian layer, with a supra-crustal Sinian–Jurassic sedimentary cover (Wu *et al.*, 2001). The Huangling anticline is a major NNE–SSW-oriented structure in the area, about 50 km in length, to the south-east of Zigui, and its core is composed of pre-Sinian metamorphic and magmatic rocks (Figure 21.11). The strength and stability of this anticlinal structure are the principal reasons for the siting of the Three Gorges Dam in its location, near Sandouping town (about 60 km west of Yichang city).

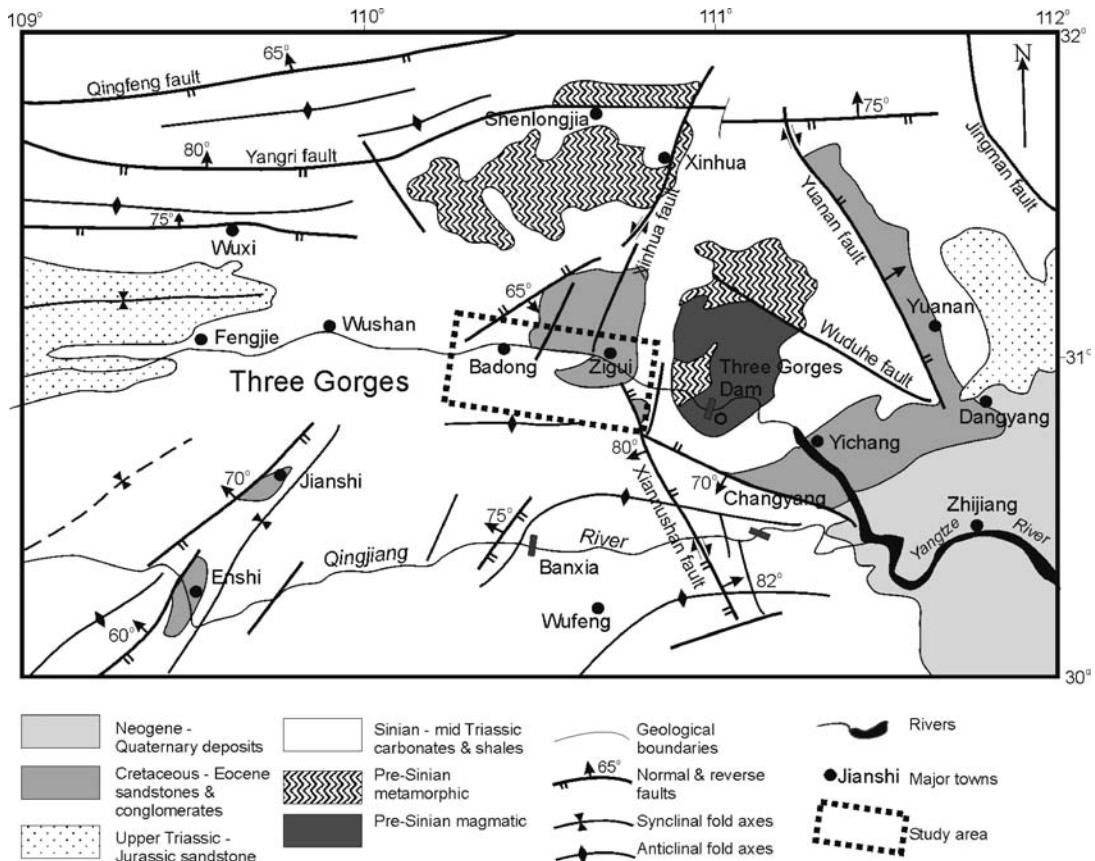
There are three main fault and fracture systems in the area (Figure 21.11). Firstly, to the south-west of the Huangling anticline is the NNW–SSE-oriented Xiannushan fault, which consists of three, parallel shear zones (Chen, 1986). Secondly, there is the NNE–SSW-oriented Jiuxiwan fault, which intersects the Xiannushan fault, near Zigui, and disappears just to the south of it. Thirdly, the Zigui Basin is crossed by the Niukou–Gongpiao fault zone, which is of similar orientation to the Jiuxiwen fault.

In addition, the area to the south of Zigui and Badong is characterized by a system of secondary faults, which follow the orientation of the fold system in the area, i.e. ENE–WSW (Wu *et al.*, 1997). These secondary fault and fracture systems tend to form ‘weak’ zones, which favour slope instability.

#### 21.2.2.3 Land instability

A variety of slope failure types occur in this region. They include simple, rotational slumps in poorly or unconsolidated materials, translational rock and debris slides, debris flows and complex examples involving more than one type of failure mechanism and several types of material. As an example, Figure 21.12 shows the massive rock block sliding that happened on 13 July 2003 in Qian-jiang-ping, Zigui County, blocking the Qinggan River, 3 km from the river mouth discharging to the Yangtze and just about 20 km upstream of the dam.

According to our field observations and published articles in Chinese journals (China Yangtze Three Gorges Project Development Corporation, 1999; Chinese Environment Agency, 2001; Three Gorges dam and reservoir Project, 2002), the massive urban development required for the relocation of major county towns to nearby higher positions has already triggered several large landslides. For instance, the Huangtupo landslide has forced the



**Figure 21.11** Regional geological and tectonic framework map After Wu Shuren *et al.* (1997)

new town of Badong County to be relocated about 6 km to the west in Xirangpo. A landslide, on 3 March 2002, moved more than 20 million cubic metres of debris down the slope, threatening the new town of Wushan County.

### 21.2.3 Methodology: multi-variable elimination and characterization

As shown in Table 20.3, the relatively high spatial resolution in the VNIR bands (with a pixel size of 15 m), the high spectral resolution in the SWIR and multi-spectral TIR imagery and, importantly, the along-track stereo capability make ASTER an ideal data source for geological and geomorphological interpretation (Welch *et al.*, 1998; Yamaguchi *et al.*, 2001). Our investigation is based largely on Terra-1 ASTER level 1B imagery acquired in May

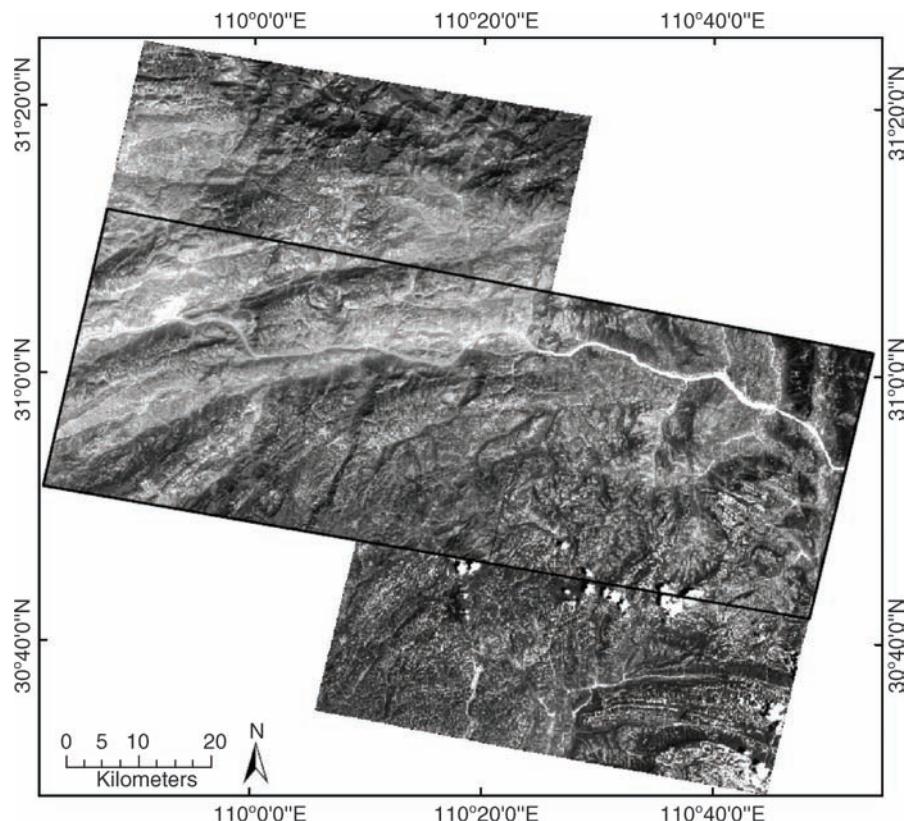
2002 and the two scenes, with a swath of  $60 \times 60 \text{ km}^2$  each, cover the whole study area from Wushan County to Badong County and to Zigui County (Figure 21.13).

There are various approaches to the generation of a landslide hazard map. Examples can be found in Mason and Rosenbaum (2002), Mantovani *et al.* (1996), Hartlen and Viberg (1988) and Varnes (1984). The data available in this study do not allow a detailed statistical assessment of the temporal and spatial distribution of landslides in the study area and therefore a semi-quantitative, logical elimination and characterization approach has been applied. Our assessment model is based on the information that can be extracted from the ASTER image data, cartographic and survey materials, published literature and limited field observations.

The method can be broadly divided into three parts: (i) parameter (relevant to slope instability) selection and model configuration; (ii) model



**Figure 21.12** A massive rock block landslide which occurred on 13 July 2003, in Qian-jiang-ping, Zigui County, blocking the Qinggan River, 3 km from the river mouth discharging into the Yangtze, and about 20 km upstream from the Three Gorges Dam



**Figure 21.13** Two Terra-1 ASTER level 1B images and extent of study area (black rectangle)

**Table 21.4** Classification and quantification of landslide-related parameters for different landslide types (R, B and D represent rockfall, block slide and shallow debris slide)

Parameter	Class	Landslide types		
		R	B	D
Lithology	Massive limestone	3	1	0
	Sandstone and shale	1	3	1
	Mudstone and debris	0	3	3
Slope angle	Steep	3	3	1
	Intermediate	1	2	3
	Gentle	0	1	1
Dissection density	High	2	2	2
	Intermediate	1.5	1.5	1.5
	Low	1	1	1
Distance from drainage	Near	1	3	3
	Distant	1	1	1
Distance from lineaments	Near	2	2	1
	Distant	1	1	1
Distance from lithological contacts	Near	2	3	2
	Distant	1	1	1

implementation: thematic information extraction and multi-data layer generation; and (iii) landslide hazard index computation and mapping.

#### 21.2.3.1 Geometric mean model configuration for landslide susceptibility mapping

A general model has been established on the basis of simplified geology and geomorphology and field knowledge of factors relevant to slope stability, shown in Table 21.4. The geometric mean serves to achieve multi-variable elimination and characterization in the model for the susceptibility mapping of three major slope failures: block slide, shallow debris slide and rockfall.

The geometric mean is defined as

$$GM = \left( \prod_{i=1}^n P_i \right)^{\frac{1}{n}} \quad (21.2)$$

where  $P_i$  is the quantization value for factor  $i$ .

The geometric mean, based on multiplication, is fundamentally different from the WLC (Weighted factors in Linear Combination) approach that is based on an arithmetic mean. A simple WLC resembles a parallel connection system, which

allows all members to survive from the beginning to the end. The geometric mean method represents a sequential connection system, which terminates whenever a zero value occurs, and is therefore effective in eliminating irrelevant areas. For instance, a zero value for a ‘flat’ area (where landslides are considered unlikely) will produce a geometric mean of 0, regardless of the values of other factors, and all ‘flat’ areas will be eliminated by the system. The geometric mean is of a ‘harsh’ nature and conceptually Boolean and, therefore, the expected result is a discrete characterization rather than a quantitative measure of regional hazard levels.

The quantization level of a variable in this model behaves like weighting in the WLC approach (see also Section 18.5.5). A high quantization value will make a variable more influential than the one with a low quantization value. For this reason, variables in the model were rescaled to a value range representative of their importance. Each parameter is ranked in no more than four levels, between 0 and 3. The quantization in these simple levels enables the geometric mean model to perform a logical elimination operation, while maintaining characteristic numerical values for final assessment.

The angular relationship between bedding attitude of sedimentary rock formation, slope aspect and slope angle is a very important factor in slope stability (Meentemeyer and Moody, 2000; Wen *et al.*, 2004). For instance, a slope that is in the same aspect as the rock beddings but has a steeper slope angle than the dipping angle of the beddings is more prone to landslide than a slope against the bedding aspect. However, accurate data of this relationship can only be collected from detailed field surveys. We have to accept the limitation of a study based on remote sensing imagery to exclude this parameter from the model described above. Climatic data are also not included because the limited study area can be safely considered under the same climate. Weather conditions are important for predicting the temporal characteristics of the landslides in a high-risk area; for instance, most landslides are triggered by heavy rains. However, our research on the spatial distribution of high-risk areas of landslide hazard is aiming to answer the question of where the hazard is likely to happen under similar weather conditions rather than when.

### 21.2.3.2 WLC model for landslide hazard assessment based on susceptibility mapping

Landslide hazard can be defined as the probability of occurrence of a landslide event of a given size, and can be estimated as the product of susceptibility, frequency and magnitude (Lee and Jones, 2004):

$$\text{Hazard} = \text{Susceptibility} \times \text{Frequency} \times \text{Magnitude}.$$

Thus, for the given three types of slope failures, the total landslide hazard level for the study area can be estimated as a linear combination of susceptibility, frequency of occurrence and magnitude of these types as follows:

$$H_{\text{Landslide}} = H_R + H_B + H_D = w_R S_R + w_B S_B + w_D S_D \quad (21.3)$$

where  $H$  represents hazard index and  $S$  susceptibility;  $w$  = Frequency  $\times$  Magnitude is weighting; and R, B, D denote rockfall, block slide and debris slide.

The frequency and magnitude data of each type of land instability have to be collected through intensive fieldwork but this work can be made very

efficient when guided by the susceptibility maps produced based on the geometric mean model from remote sensing image data.

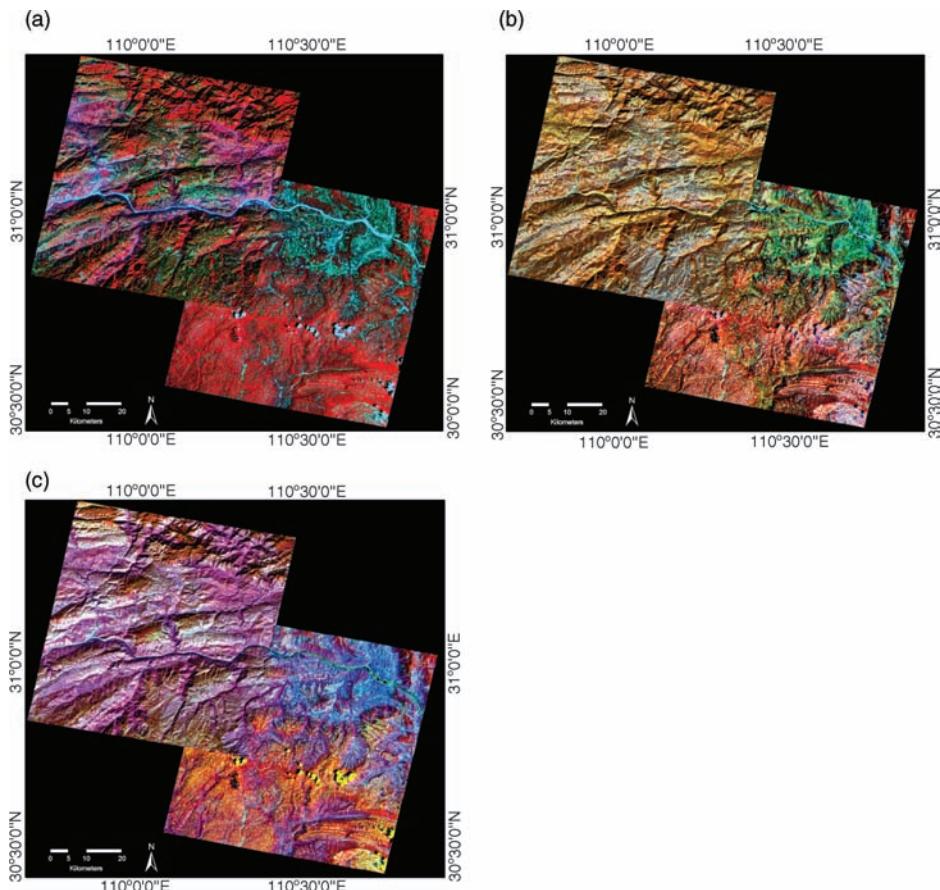
### 21.2.4 Terrestrial information extraction

#### 21.2.4.1 Map of lithological stability

Lithology is one of the most relevant parameters in landslide hazard. Different lithologies respond differently to erosion agents and conduct mass movement under differing natural conditions. For the purpose of landslide hazard assessment, a simple lithological map showing the broad categories of rock types representing high, intermediate and low competence is adequate. The essential information of these three lithological units can be extracted through enhancement and interpretation of the ASTER multi-spectral imagery. A series of simple colour composite images, enhanced using DDS with the following band combinations, were used:

- *Bands 3–2–1 RGB*: (Figure 21.14a): This image provides detailed information of textures relating to topography, geomorphology and geological structure and vegetation (see Table 20.3 for the spectral ranges of ASTER image bands).
- *Bands 4–6–9 RGB*: (Figure 21.14b): In this image, discrimination between the limestone–shale units, and the interbedded pelites and psammites, is clear.
- *Bands 4–6–12 RGB*: (Figure 21.14c): This image highlights sandstones because of the strong thermal emissivity of quartz, revealed in band 12 (8.925–9.275  $\mu\text{m}$ ).

The lithological information derived from visual interpretation of these three images, integrated with information from the published small-scale geological map, was the basis for mapping the three broad lithological units with low, intermediate and high competence as listed in Table 21.5 and shown in Figure 21.15. In the Three Gorges, different types of slope failures and different degrees of landslide susceptibility are associated with different lithologies. Soft lithologies like Quaternary deposits and mudstone may give rise to shallow debris slides, while massive limestones tend to form stable ridges and high peaks that can be subject to rockfalls. Accordingly, the three lithological units were given



**Figure 21.14** ASTER colour composite images for lithological mapping: (a) bands 3–2–1 in RGB; (b) bands 4–6–9 in RGB; and (c) bands 4–6–12 in RGB

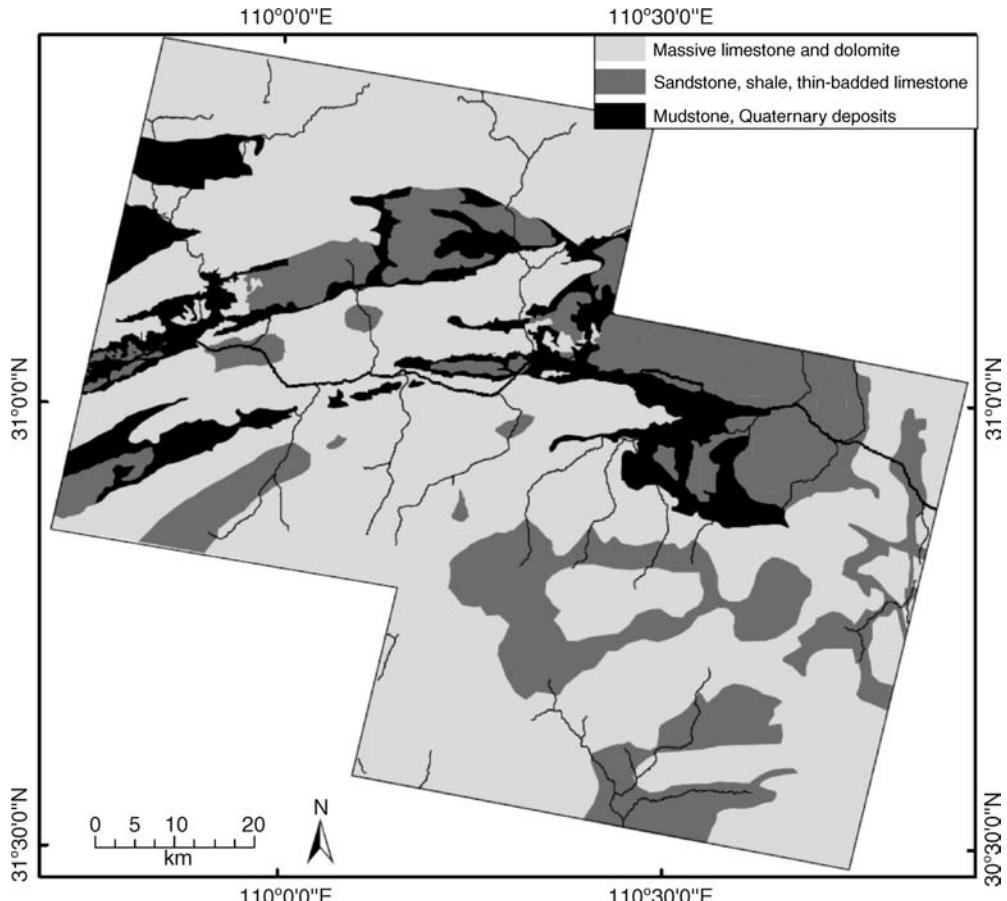
different rankings for different types of slope failures (Table 21.5).

The assignment of ranking values in Table 21.5 is not intended to provide a precise quantitative estimate of the rock stability based on a cardinal scale,

but to establish an ordinal ranking between classes of lithologies in relation to the likelihood of particular slope failure. The ranking values for lithology are 0, 1 and 3 instead of consecutive integers 0, 1 and 2. This is to emphasize the influence of the lithology

**Table 21.5** Lithological competence classes and their stability ranking for rockfall (R), block slide (B) and shallow debris slide (D)

Lithological classes	Competence	Stability ranks		
		R	B	D
Massive limestone and dolomite	High	3	0	0
Sandstones, thinly bedded limestone and shale	Intermediate	1	3	1
Shale, mudstone and sandstone associations and Quaternary deposits	Low	0	3	3



**Figure 21.15** Lithological map of the study area classified on the basis of the competence of rock formations

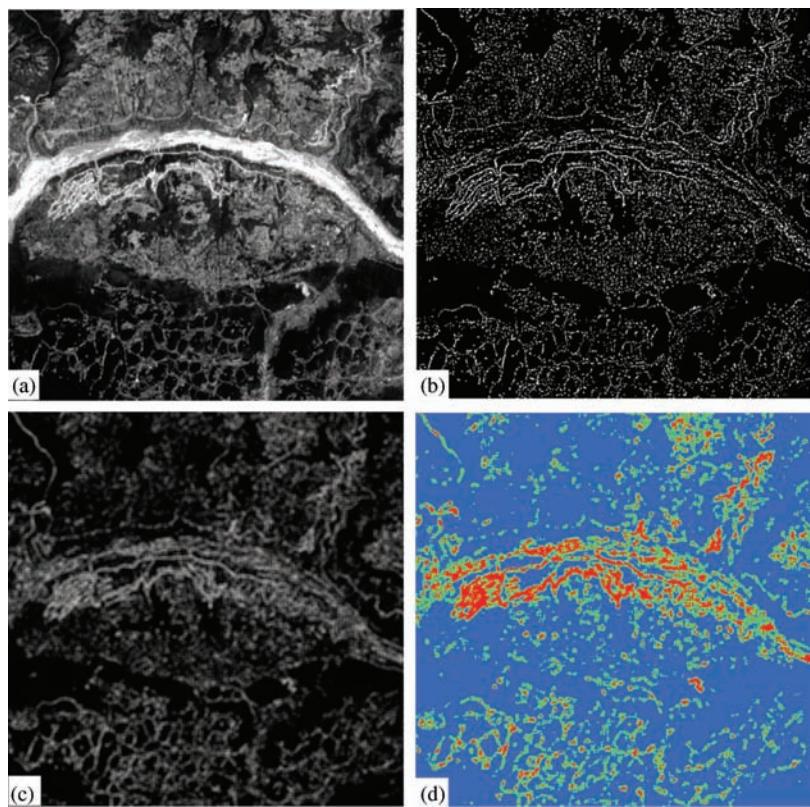
in the geometric mean mode, which reflects our field observation of the relationship between different rock competence and type of slope failure.

#### 21.2.4.2 Map of dissection density

Intensely dissected terrain contains many natural pathways for surface runoff, which exacerbates erosion along steep gullies, and is more vulnerable to landslides than less dissected land surfaces. The image representing dissection density is a measure of ground surface textural complexity. As illustrated in Figure 21.16 for the Badong area, a texture density image can be derived through a simple and effective sequence of edge enhancement (Laplacian filtering), thresholding (Figure 21.16b) and smoothing (Figure 21.16c). ASTER band 2 (red band) in Figure 21.16a was chosen for this purpose because of its 15 m spatial resolution and minimal effect of

vegetation at the wavelength. The map of dissection density (Figure 21.16d) is then produced from the texture density image by thresholding based on the natural breaks of its histogram corresponding to broad classes of low, intermediate and high dissection density (Table 21.6). Here, instead of using consecutive integer numbers, 1, 1.5 and 2 are used to reduce slightly the significance of this parameter. Figure 21.17 is the finally derived dissection density map of the study area.

From Figure 21.17, it appears that urban development (roads and buildings) increases the dissection density dramatically. This phenomenon has no relationship with the dissection of natural land surface but is relevant to the land stability assessment because 'human erosion' activities have destabilized those marginally balanced slopes and triggered mobilization of old settled landslides. It



**Figure 21.16** Derivation of dissection density map. (a) ASTER band 2 image. (b) The texture image produced using a Laplacian filter and thresholding to exclude negative values. (c) Smoothed texture image using a  $7 \times 7$  smoothing kernel; the DN values in this image give a measure of texture density. (d) The dissection density map produced by ‘slicing’ thresholding of image (c): red, highly dissected; green, intermediate dissection; blue, least dissected

is reasonable to consider natural and anthropogenic dissection together as one parameter for regional assessment in this rapidly developing area.

#### 21.2.4.3 Buffer map of distance from faults and lineaments

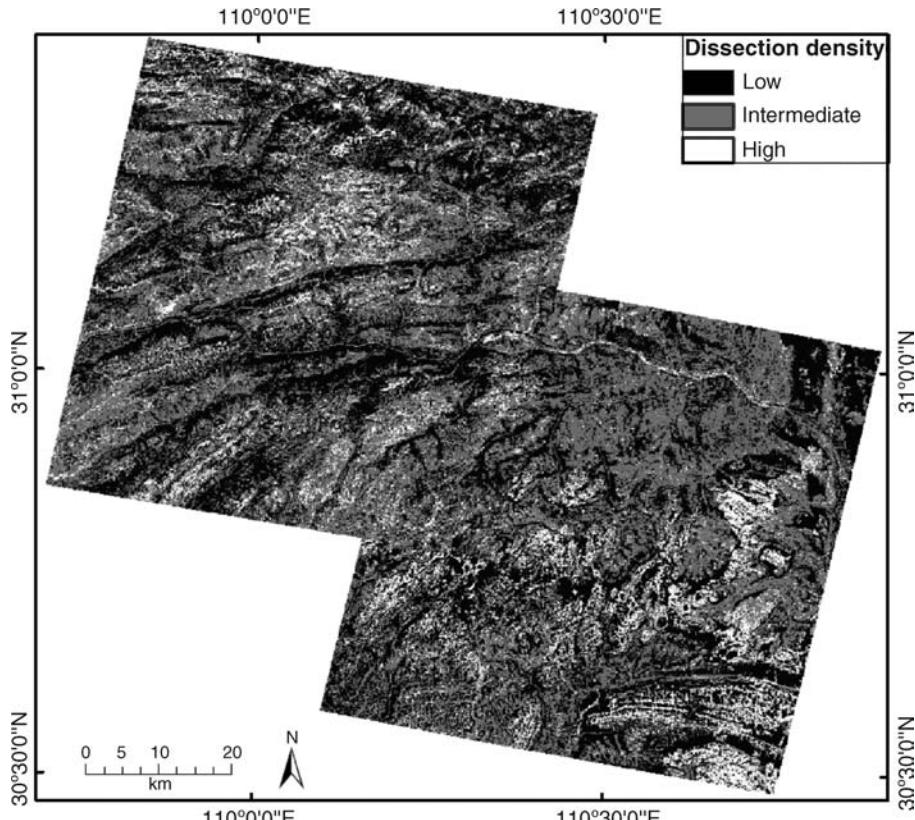
Geological structures such as faults and fractures form discontinuities in rock formations. On small

scales, such discontinuities contribute to the complex textures. On a larger scale, neo-tectonic movements, such as earthquakes, often produce ‘broken zones’ along large fault segments in recent superficial deposits, and can be the trigger for landslides. These weak zones present favourable conditions for landslides (Saha, Gupta and Arora, 2002). We therefore include major structural discontinuities, namely faults and fractures, as a parameter in the analysis of land stability. A map of faults and fractures shown in Figure 21.18 was produced based on the interpretation of ASTER imagery in combination with published maps.

Field evidence suggests that lineaments of greater extent can have a greater influence on slope instability than smaller features. The influence that lineaments exert on slope instability was thus estimated through the definition of buffer zones using a

**Table 21.6** Dissection density classes

Class	Dissection density	Values
Low texture density	Low	1
Average texture density	Intermediate	1.5
High texture density	High	2



**Figure 21.17** Dissection density map of the study area

distance function relating to lineament length. Two classes of ‘Near’ and ‘Distant’ were then defined (Table 21.7). Buffering creates vector polygons around vector linear features at a specified distance; the buffer polygons can then be converted into a raster dataset for map overlay calculations.

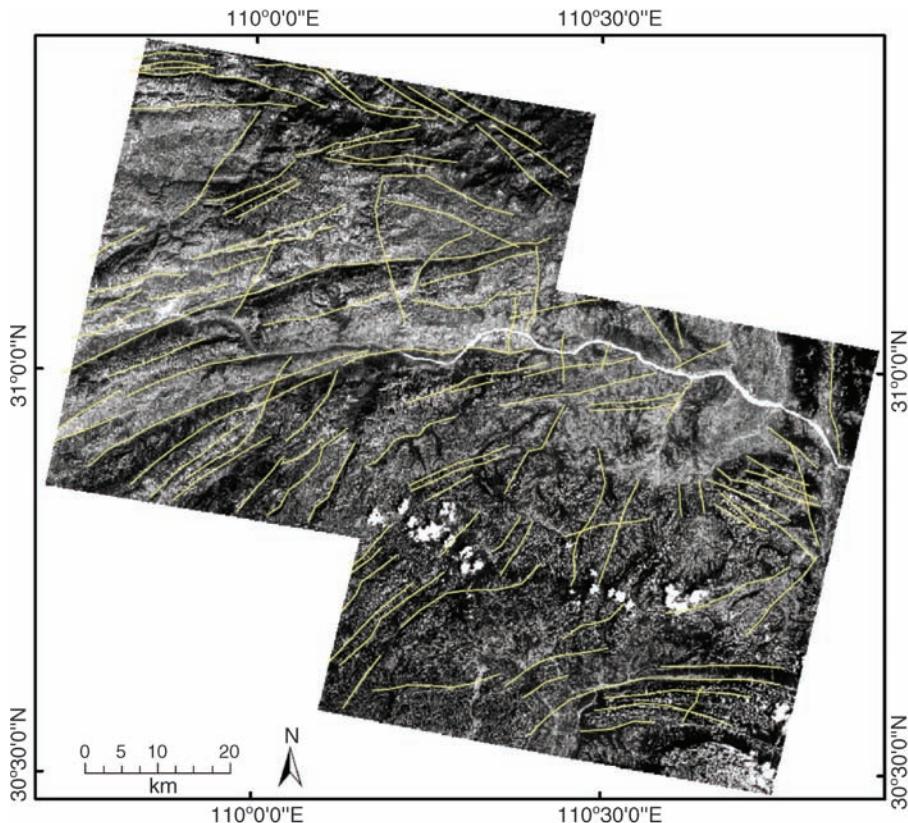
#### 21.2.4.4 Buffer map of distance from lithological contacts

The inter-layering of formations of contrasting physical properties, such as strength and permeability, can lead to differential reaction to changes in the environment, such as increase in pore-water pressure and reduction of shear strength; this behaviour has been found to promote slope instability, particularly of the translational block-sliding type (Guzzetti *et al.*, 2003).

In the Three Gorges, high terrain susceptibility to block sliding was observed in stratigraphic contact between limestone and mixed layers (mudstone,

sandstone and shale). In the case of mixed layers overlying limestone, the limestone could provide sliding surfaces for the overlying strata to slide upon. In the case of limestone overlying mixed layers, water may be concentrated in limestone due to its high secondary permeability, which may result in springs along the boundary of underlying clay-rich impermeable lithologies; this condition saturates the underlying softer layers and promotes instability (Hutchinson, 1995).

The lithological stability map (Figure 21.15) was used to delineate the contacts between lithologies of contrasting physical properties. Mudstone, sandstone and shale formulate the ‘mixed layers’ category, whereas the ‘limestone’ category comprises both thinly bedded and massive formations. With the contacts between groups of contrasting competence identified, buffer zones 100 m wide were generated along these contacts to produce the map shown in Figure 21.19 (see Table 21.8).



**Figure 21.18** Major faults and lineaments used in the generation of the fault buffers, displayed over the ASTER band 2 image

### 21.2.5 DEM and topographic information extraction

A DEM with a spatial resolution of 45 m was generated from the ASTER images used in this work. The DEM conforms to WGS84 datum and UTM N49 projection. The essential topographical

information including slope angle and drainage was extracted from the DEM data.

#### 21.2.5.1 Slope map

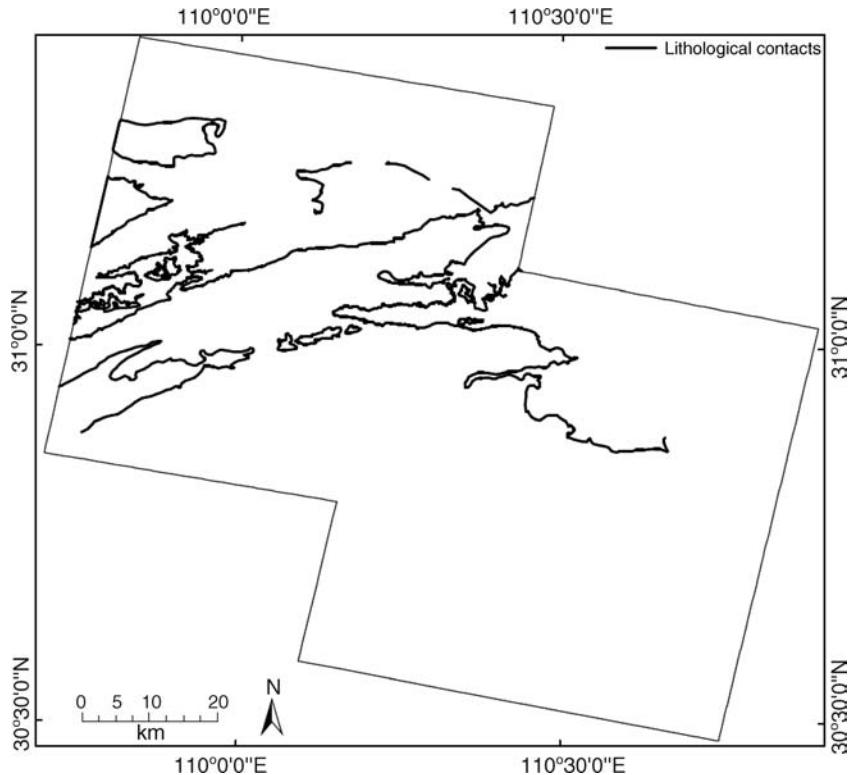
Slope angle and geometry are controlling factors in slope stability in the Three Gorges Reservoir Region (Huang and Li, 1992; Wu *et al.*, 2001), and a digital slope image is therefore a fundamental part of landslide hazard assessment. Slope angle was extracted from the DEM, using a  $3 \times 3$  calculation kernel based on the formula (4.15) in Part One of this book;

$$\tan G = \sqrt{\left(\frac{\partial z}{\partial x}\right)^2 + \left(\frac{\partial z}{\partial y}\right)^2} = \sqrt{g_x^2 + g_y^2}$$

where  $G$  is the slope angle,  $\partial z$  the  $z$  elevation increment,  $\partial x$  and  $\partial y$  are the horizontal increments

**Table 21.7** Distances from faults and lineaments

	Lineament buffer distance (m)		
Lineament length (m)	<2500	2500–5000	>5000
Near	0–100	0–200	0–500
Distant	>100	>200	>500



**Figure 21.19** Map of lithological contacts between formation of contrasting physical properties

in column and line directions of the DEM. For raster data, both  $\partial_x$  and  $\partial_y$  are equal to 1. Thus the calculation kernel is

$$g_x = \begin{pmatrix} 0 & -1 & 1 \end{pmatrix} \quad g_y = \begin{pmatrix} 0 \\ -1 \\ 1 \end{pmatrix}.$$

Field evidence suggests that different lithologies have different critical slope angles for slope failure.

**Table 21.8** Distance from lithological contacts classes

Lithological setting	Distance from lithological contacts (m)		
	Limestone over mixed layers	Mixed layers over limestone	
Near	0–100	0–100	
Distant	>100	>100	

A competent rock formation usually has a higher critical slope angle than that for a soft lithological formation. Field measurements were used to estimate critical slope angles for different lithologies and divide the slope map into lithology-controlled stability classes (Table 21.9). That led to a classification scheme whereby, for the same slope class in terms of equivalent stability, instead of the same slope angle, the threshold angle for a slope class varies depending on the competence of the lithology. Slopes in mudstones and Quaternary deposits with angles below 5° are largely stable but become unstable when the slope is greater than 15°. Slopes composed of sandstone, shale and thinly bedded limestone are largely stable below 15° slope angle and unstable above 25°. Massive limestone, the most competent lithology in the area, is usually not subject to slope failure below 20°, although a slope angle above 35° could be subject to rockfalls and topples. The slope angle map was thus classified into relative categories of 'Gentle', 'Intermediate' and 'Steep' slopes based on the competence of the lithology (Figure 21.20).

**Table 21.9** Slope angle class intervals chosen by competence of lithology

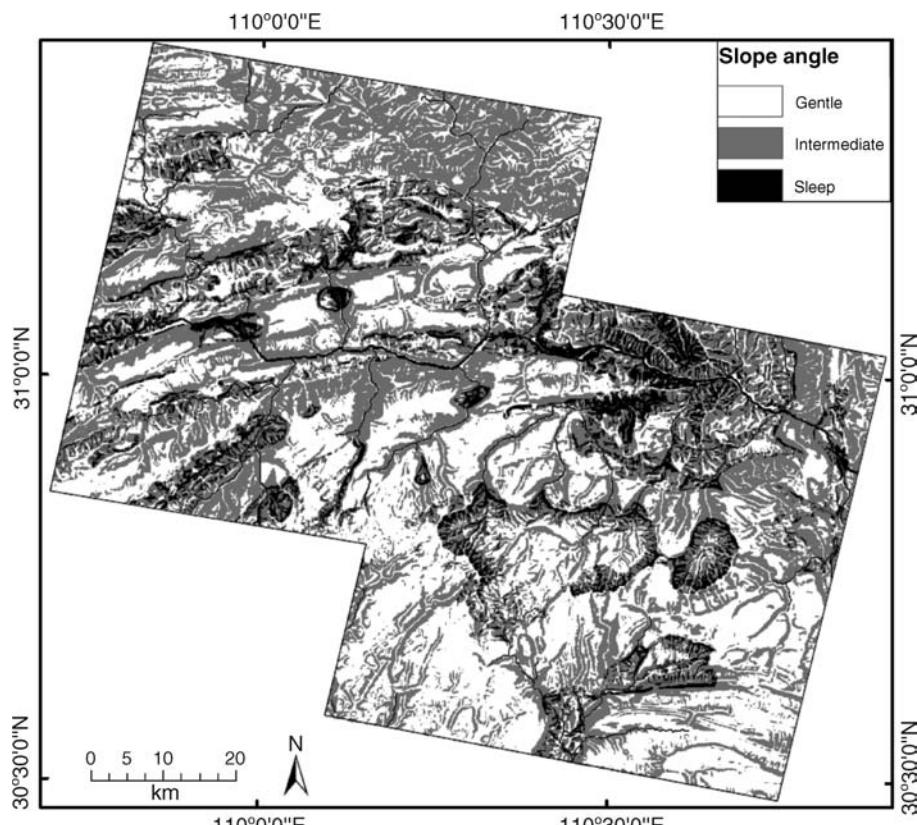
Class	Slope angle range (degrees)			Coverage (%)
	Limestone	Sandstone	Mudstone	
Gentle	0–15	0–10	0–5	43
Intermediate	15–35	10–25	5–15	48
Steep	>35	>25	>15	9

### 21.2.5.2 Buffer map of distance to drainage

Many of the large landslides in the Three Gorges Region occur in close proximity to water courses. The erosion process at the foot of the river banks destabilizes the slope and leads to landslides. River terraces (composed of alluvium, sand and gravels) are also prone to collapse during heavy rainfall. The distance from rivers is therefore considered an important factor in characterizing slope stability.

In general, larger drainage channels can have a greater influence upon slope instability than smaller ones. This variation of influence of drainage network to slope instability can be captured through the distance buffer zones of different widths.

The drainage network of the study area was automatically extracted from the DEM using RiverTools software and then classified into three broad categories: minor streams, tributaries to the Yangtze and the Yangtze River itself for which



**Figure 21.20** Map of slope angle classified on the basis of the competence of underlying lithology

**Table 21.10** Distance buffer from drainage network

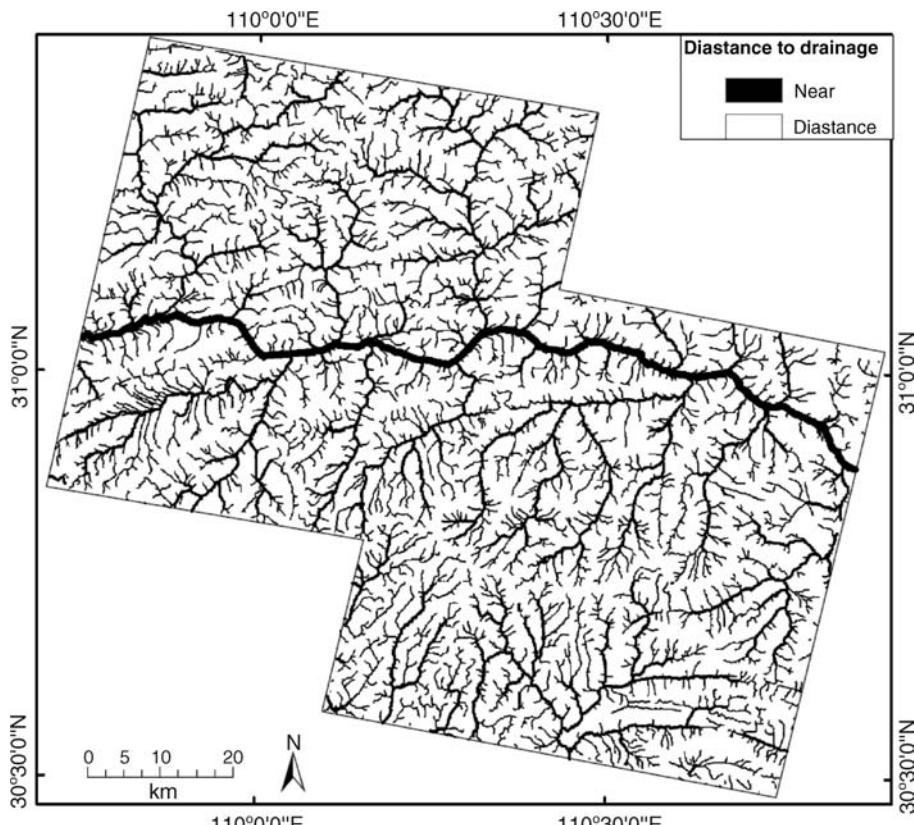
Stream type	Drainage network buffer distance (m)		
	Yangtze River	Major tributaries	Minor streams
Near	0–500	0–200	0–100
Distant	>500	>200	>100

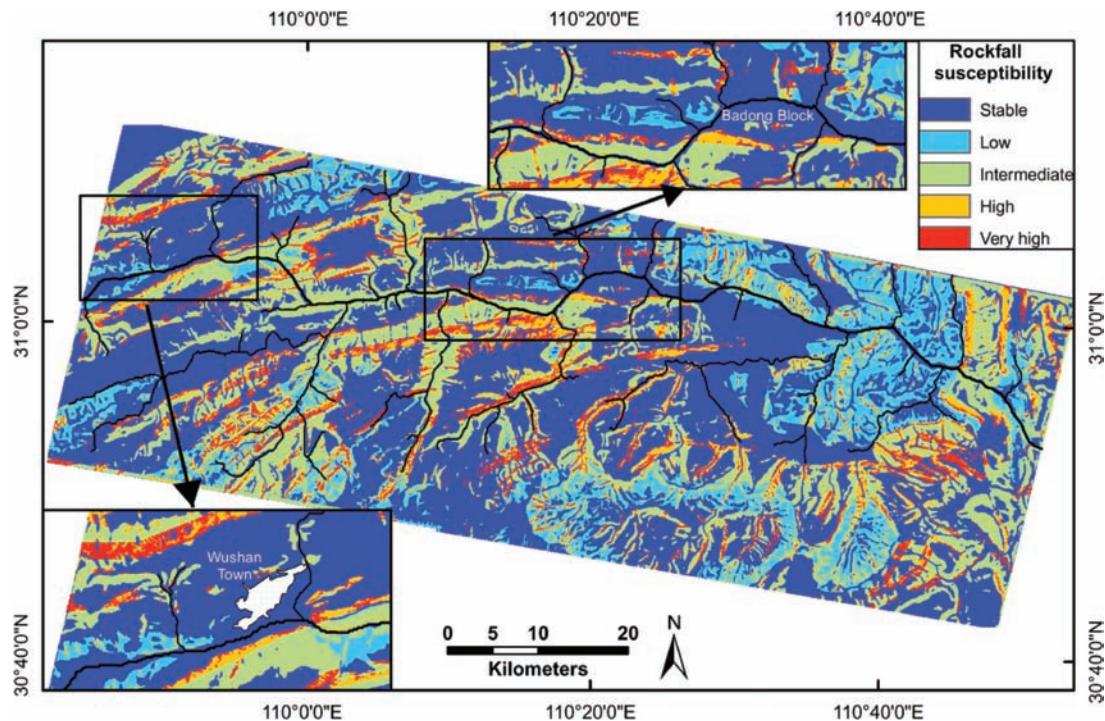
buffer distances of 100 m, 200 m and 500 m were assigned (Table 21.10). A buffer map of distance to drainage was then produced (Figure 21.21).

### 21.2.6 Landslide hazard mapping

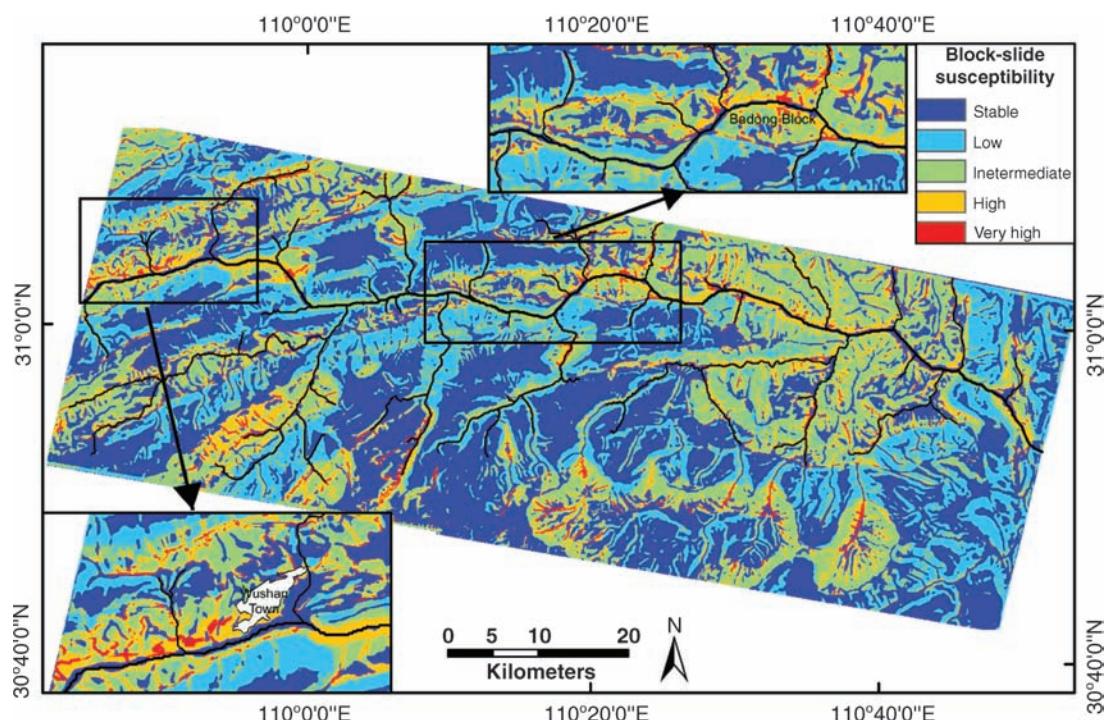
With six essential parameters represented as raster image layers containing simple ranking numbers for three different slope failure types specified in Table 21.4, the geometric means were calculated for rockfall, block slide and shallow debris slide. The

values of the geometric mean for each slope failure type were then classified into five classes, from stable to high susceptibility, based on histogram breaks, and thus maps of susceptibility to rockfalls, block slides and shallow debris slides were produced. The spatial distribution of susceptibility is different for the three failure modes. While steep slopes in limestone mountains are more susceptible to rockfalls (Figure 21.22), valley slopes formed in sandstone and shale sequences appear to be prone to block slides (Figure 21.23). In areas of soft

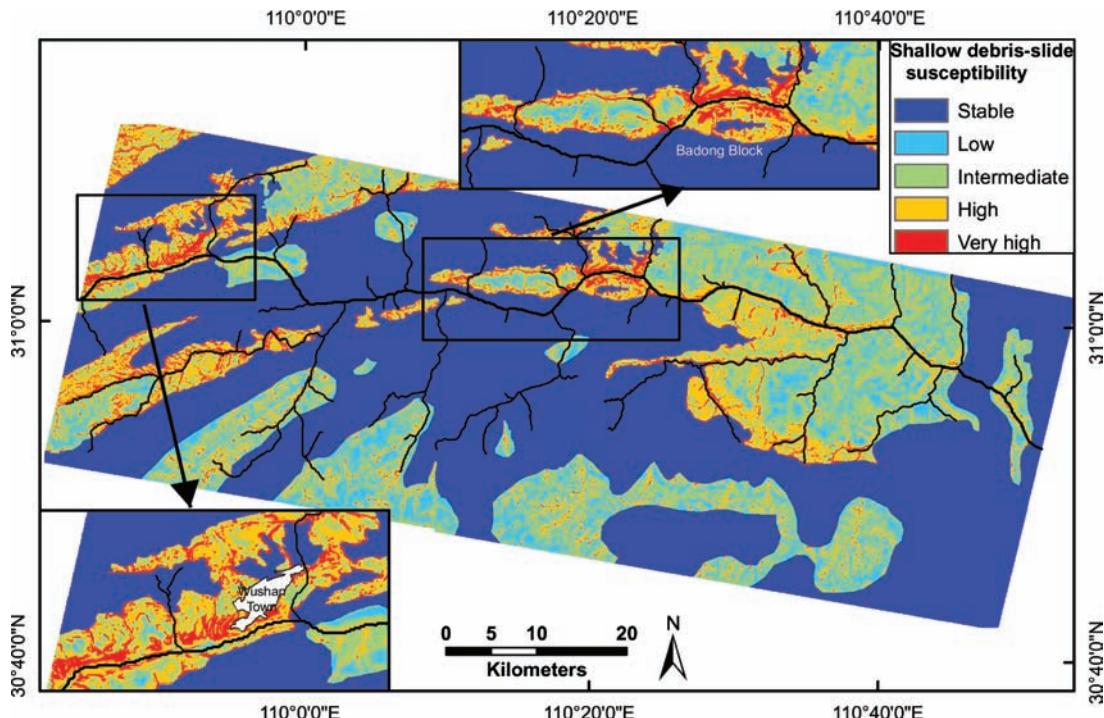
**Figure 21.21** Drainage network extracted from the DEM and buffered at 100, 200 and 500 m



**Figure 21.22** Susceptibility to rockfalls in Wushan–Badong–Zigui



**Figure 21.23** Susceptibility to block slides in Wushan–Badong–Zigui



**Figure 21.24** Susceptibility to shallow debris slides in Wushan-Badong-Zigui

lithologies, such as mudstone and Quaternary deposits, shallow debris slides are the dominant type of land instability (Figure 21.24). The new town areas of Wushan and Badong are subject to a high risk of this type of landslide as the urban development has to choose flat areas of soft lithology and the engineering work, such as road cutting, intensified the dissection intensity of the land surface.

The three susceptibility maps were then combined using the WLC model specified in formula (21.3) to generate the final landslide hazard map as shown in Figure 21.25. In the calculation of the landslide hazard index, the weights for each map in the WLC model are based on the qualitative rating of frequency and magnitude of different landslide types estimated from field observation data and published landslide data as shown in Table 21.11.

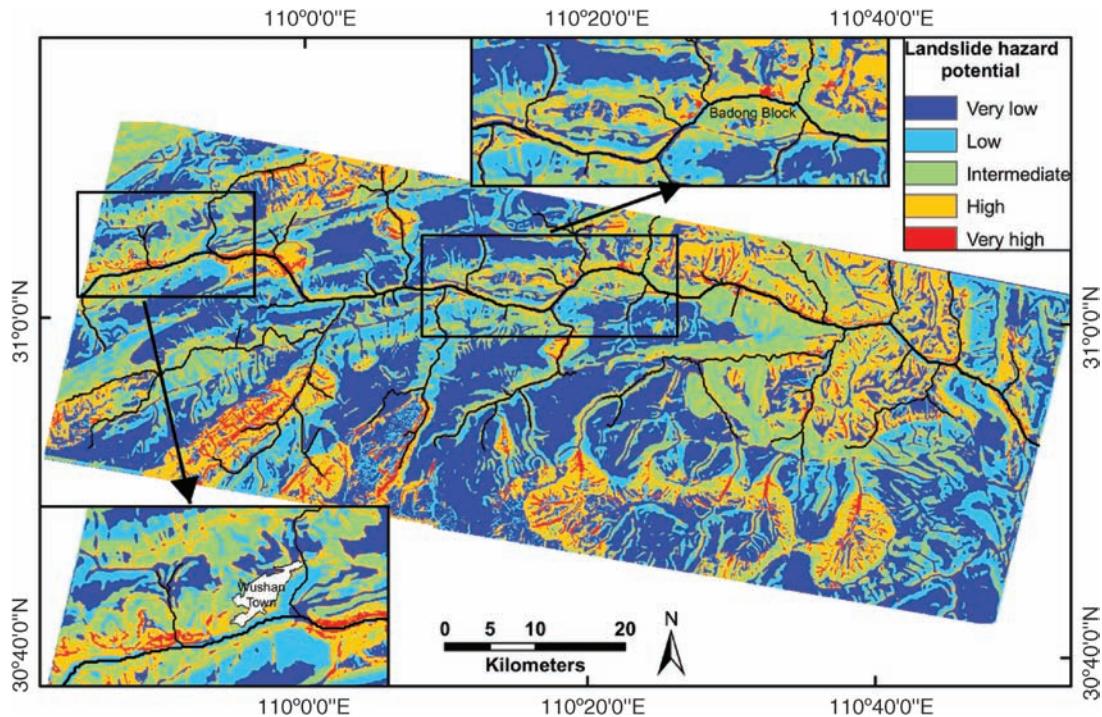
### 21.2.7 Summary

Despite the paucity of data and the simplicity of the model based on ASTER satellite imagery, this study

has shown convincing success in the delineation of areas most susceptible to landslide hazard in the Three Gorges Region according to our field investigation, collected data of known major landslides and information of ongoing engineering work to fix unstable slopes in the region. The results indicate that the main areas of concern are connected very closely to the recent relocation and development of new towns in the Three Gorges Region.

The mathematical nature of the proposed landslide hazard index is effective in masking off the stable areas; attention can thus be focused on the areas subject to high levels of hazard and provides effective guidance for further investigations and recommendations for engineering measures.

This region is clearly subject to widespread slope instability, irrespective of anthropogenic influences, but recent construction activities have triggered and reactivated several large landslides. Reservoir flooding and the consequent raising of the shoreline have the potential to change and rejuvenate slope profiles, and to trigger new landslides. These in turn may affect reservoir capacity and dam safety. This work of regional assessment of



**Figure 21.25** Landslide hazard map of Wushan–Badong–Zigu

**Table 21.11** Qualitative rating of frequency and magnitude of the three types of slope failure in the Three Gorges Region

Landslide type	Frequency	Rating	Magnitude	Rating	Weighting
Rockfall	Low	2	Medium	2	4
Block slide	Medium	3	High	3	9
Shallow debris slide	High	4	Low	1	4

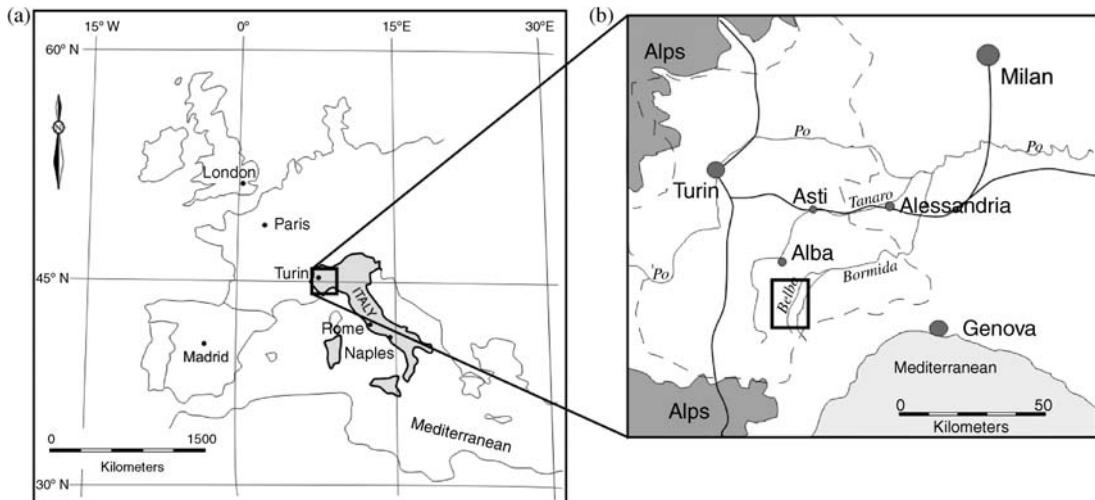
landslide hazard highlights the most vulnerable areas and the need for immediate and long-term action plans to ensure that further developments proceed within suitable engineering guidelines.

The work has demonstrated that ASTER imagery is a very useful source of topographic and spectral information for regional landslide hazard mapping. The 14 multi-spectral bands (in the VNIR, SWIR and Thermal IR) of ASTER and its stereo capability facilitate mapping and assessment of landslide hazard on a regional scale and especially where detailed geological maps and topographic maps are not available.

## 21.3 Predicting landslides using fuzzy geohazard mapping; an example from Piemonte, North-west Italy

### 21.3.1 Introduction

This section describes the use of multi-form, digital image data, within a GIS-based multi-criteria analysis model, for the regional assessment of risk concerning slope instability (Mason *et al.*, 1996; Mason, Rosenbaum and Moore, 2000; Mason and



**Figure 21.26** (a) Map of western Europe, showing the location of Piemonte; and (b) the Langhe Hills in Piemonte, NW Italy (the dashed line indicates the regional administrative boundary and the small bold box indicates the study area)

Rosenbaum, 2002). Landslides are documented throughout the Piemonte region of NW Italy but they are a significant problem in the area known as the ‘Langhe’, a range of hills south-east of Turin (Figure 21.26), where slope instabilities have been experienced over a prolonged period. An exceptional storm event in November 1994 produced extensive flooding and widespread mass movements, leading to many fatalities and consequential damage to property. The response to all natural disasters demands an assessment of the hazard and some prediction of the likelihood of future such events. A recent and full National Landslide Inventory has shown that more than 34 000 landslides affect the Piemonte region in different geological contexts. Recent research into the understanding and quantification of the problem in Piemonte has also included the creation of a permanent scatterer system for monitoring slope movements using InSAR (Meisina *et al.*, 2008).

Satellite remote sensing enables rapid and routine collection of data over a much greater area than can be obtained from a typical ground-based survey of comparable cost. In cases where landslides could occur across a sizeable region, remote sensing may be the only readily available source of information concerning the terrain, particularly in the aftermath of a devastating event or where erosion could rapidly remove the evidence. Remote sensing can provide

information on the surface morphology (arcuate scarp, hummocky ground, tension cracks and disrupted drainage), changes in vegetation as a result of increased water content, and soils which contain a lot of water (i.e. are poorly drained). The remotely sensed information is supported by field measurements using reflectance spectroscopy and X-ray diffraction (XRD) which provides direct information concerning the local soil mineralogy. Oxidized iron is an indicator of intense leaching and weathering of iron-bearing rocks and gives a very distinctive signature in remotely sensed imagery as well as colour to soils. Intensely weathered and fractured rocks are zones of inherent weakness that may indicate locations where mass movement is likely to be initiated. Clays and iron oxides have therefore been employed as the two main soil mineralogical targets within this investigation, to establish their association with landslide occurrence with a view to determining their utility as geohazard indicators for mass movement on a regional scale.

An important aspect of this study was, therefore, to identify the temporal and spatial distribution of areas liable to movement, including the location of potential slip surfaces. This study considered the geomorphological and mineralogical expressions of mass movements, in addition to some engineering considerations, with a view to producing the geohazard assessment.

### 21.3.2 The study area

The Langhe Hills of Piemonte (Figure 21.26) lie on the flank of the southernmost arc of the western Alps, on the margins of the plain of the River Po, and comprise a series of gently dipping ( $7\text{--}12^\circ$ ) Tertiary sediments of Oligocene (Aquitanian) age (about 26 million years). The gently north-westerly dipping strata produce a series of NE–SW trending asymmetric valleys with SE-facing, gently dipping slopes and NW-facing steep scarp slopes.

Fine-grained argillaceous rocks, such as claystone, mudstone, siltstone and shale, dominate this region and usually occur as alternating sequences of porous sandstones and impermeable mudrocks. Stratigraphy of this nature is particularly prone to differential weathering, swelling and erosion.

The area has also been isostatically active since glacial times and the geomorphology of the upper Langhe River basins suggests that the area has undergone significant Quaternary uplift and rotation (Biancotti, 1981; Embleton, 1984). This has caused a marked change in drainage characteristics, including river capture away from the Cuneo Plain north-east into the Alessandria Plain (which is several hundred metres lower).

#### 21.3.2.1 History of slope instability

Major slope movements in the area have been documented over the last hundred years or so, by Sacco (1903), Boni (1941), Cortemiglia and Terranova (1969), Govi (1974) and Tropeano (1989). Govi and Sorzana (1982) were the first to draw attention to the similarities between the various landslides in the region, noting that a close relationship existed between the timing of the landslides and the period of antecedent rainfall. They observed that some failures occurred on slopes which had been affected by similar failures in the past. They also inferred that human actions, such as the construction of road cuts, terraces, alteration of natural drainage systems and dumping of waste into fissures, can be significant factors for initiating slope instability.

One interesting aspect of this case is that the Langhe experiences very heavy rain each winter, yet the literature suggests a periodicity to major landslide events (with past events in 1941, 1948, 1972, 1974 and 1994). The map shown in Figure 21.27

illustrates the distribution of landslides produced during the last three major landslide events. This map suggests that much of the area has experienced landsliding since 1972.

Between 4 and 6 November 1994, during a severe cyclonic weather event, several hundred millimetres of rain fell on Piemonte. Many of the slope movements began with the ground cracking and bulging, but the main displacements did not start until around midnight on the 5th, coming to rest in the early hours of Sunday, the 6th. The average rainfall during each day of the 1994 storm was 33 mm, contrasting with the average monthly rainfall of around 140 mm for November in Piemonte (Polloni *et al.*, 1996). In fact, between 200 and 300 mm of rain fell between 2 and 6 November, with 90% of this falling on the 5th. On the 6th, the region received the greatest recorded rainfall in 80 years (up to 25 mm/hour).

Groundwater storage capacities of the river basins were exceeded and the water table reached surface levels; subsequently rainfall could only escape by flowing overland, causing widespread flooding. In total, 70 people were killed, several thousand people were rendered homeless, 200 settlements (towns and villages) were affected and over 100 bridges were damaged or destroyed. The total damage was estimated at approximately US \$10 billion, within an area comprising about 30% of the region (Polloni *et al.*, 1996).

#### 21.3.2.2 Slope movement types and geotechnical considerations

Two broad types of failures are observed in the region: debris flows and block slides (single surface and compound, and both first-time and reactivated). These are illustrated in Figure 21.28 and their characteristics described in Table 21.12.

Previous research (Bandis *et al.*, 1996; Polloni *et al.*, 1996; Forlati *et al.*, 1996) indicated a number of conditioning factors and situations. From their results the following conditions have been identified as being significant in triggering slope failures in this case:

- The rocks and soils are in a severely weakened state at the time of failure.
- Antecedent rainfall was critical in the initiation of debris flows.

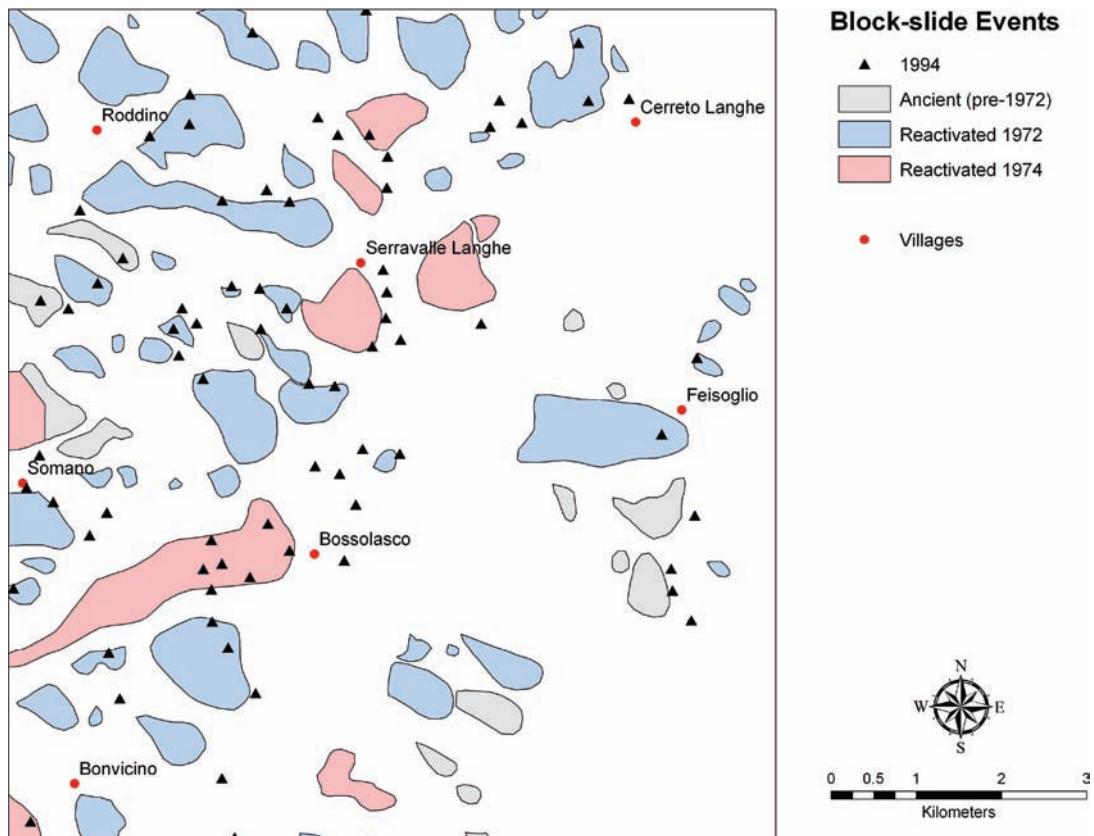


Figure 21.27 Some of the previous block-slide event localities in the Langhe region

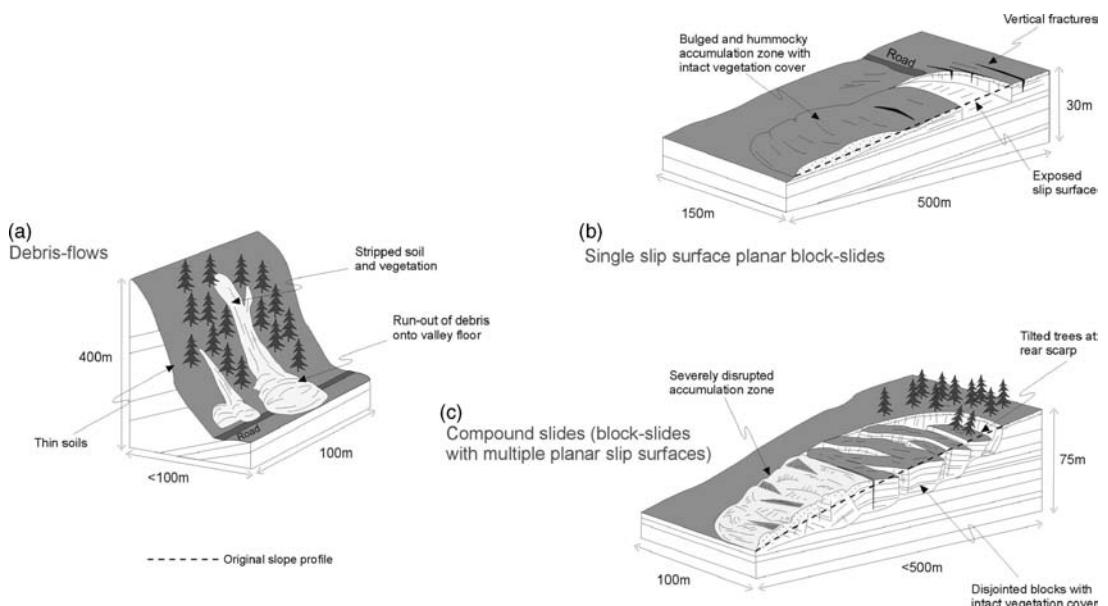


Figure 21.28 Slope movement types observed in the area

**Table 21.12** Characteristics of slope movement types

Type	Debris flows	Block slides
Movement mechanism	Shallow sheet flows (and some slides)	Translational simple and compound block slides
Slopes	20–53° (generally 20–40°)	5–15°
Attitude	At high angles to bedding	At low angles to bedding
Width/length aspect ratio	0.05–0.3	0.3–0.5
Depth	about <1.5 m	1–10 m (simple), 20–30 m (compound)
Material involved	Top soil/regolith and vegetation	Rock, soil and vegetation
Other characteristics	Commonly related to slope concavities, drainage gullies and hollows; commonly in wooded areas	Incipient phase marked by ground swelling; open fractures and tension gashes above the crown prior to failure
Timing and rates	Occur after major rainfall events. Rapid movement (a few metres per second)	Variable movement rates (between 10 and 100 m/h)
Nature	Highly destructive	Large area of the ground unbroken (simple slides); considerable disruption to the ground surface (compound slides)

- The most frequent slope gradients where failures occurred were between 30° and 40° for debris flows and between 10° and 15° for block slides.
- The position of the groundwater level (relative to ground level) and rainfall intensity are critical to slope stability; if these are both high, then slopes may become unstable even at very low slope gradients.
- The slope failure planes (in block slides) are pervasive, between 100 and 200 m in length and occur at clay-rich layers.
- Failure plane layers contain high contents of the swelling clay, smectite (montmorillonite).

Work by Bandis *et al.* (1996) also yielded valuable geotechnical parameters for the rocks and soils in this region which were used in preparing the data for the hazard assessment.

### 21.3.3 A holistic GIS-based approach to landslide hazard assessment

#### 21.3.3.1 The source data

A digital image database was compiled for this work from a number of sources. It included a DEM,

created photogrammetrically from ASTER imagery, from which slope angle (degrees) and slope aspect (degrees, as a bearing from north) were calculated. Geological boundaries, drainage and infrastructure were digitized from paper maps of 1 : 25 000 and 1 : 50 000 scales. Multi-temporal SPOT Panchromatic and Landsat TM image data were used to locate known landslides produced by the 1994 storm event and to derive rock, soil and land-use information.

#### 21.3.3.2 Selection and data preparation

A number of significant criteria were identified, on the basis of direct field observation and published work:

1. Slope morphology: Surface parameters, slope gradient and aspect were calculated from the DEM. Fuzzy functions were used to normalize these criteria to a common scale, using control points derived from field evidence.
2. Field evidence suggests that block slides occur frequently in close proximity to roads and that debris flows tend to be channelled by first- and second-order stream-valley morphology. Two criteria images were generated to represent the

Euclidean distance from roads and from drainage features, using information extracted from satellite images and published maps.

3. Geotechnical measures: Both block slides and debris flows in the Langhe are planar failures, and as such can be treated as ‘infinite slopes’ at such a regional scale of assessment (Taylor, 1948; Skempton and DeLory, 1957; Brass, Wadge and Reading, 1991). The lack of pore pressure and shear strength information, and observations that these slope failures include both rock and soil, permit such a simplified approach rather than attempting to apply a more rigorous application of limit equilibrium methods. A version of this model was used to produce a ‘factor of safety’ image layer as follows:

$$F = \frac{\text{shear strength}}{\text{shear stress}} = \frac{c' + (\gamma - m\gamma_w)z \cos^2 \alpha \tan \phi'}{\gamma z \sin \alpha \cos \alpha} \quad (21.4)$$

where  $c'$  (effective cohesion) = 0.005 kN/m<sup>2</sup> ( $c'_{\text{res}} = 0$  kN/m<sup>2</sup>),  $\gamma$  (bulk unit weight) = 24 kN/m<sup>3</sup>,  $m$  (ratio of water table depth to failure surface depth) = 1.0,  $\gamma_w$  (unit weight of water) = 10 kN/m<sup>3</sup>,  $z$  = depth to failure surface,  $\alpha$  = slope angle and  $\phi'$  = effective friction angle.

The infinite slope equation was calculated directly, on a pixel-by-pixel basis, using the slope gradient and aspect images. The other parameters were interpolated from results presented by Bandis *et al.* (1996). Based on field evidence, clearly defined ranges in slope aspect could be defined for the block slides (240 – 020°) and debris flows (020 – 240°). Again on field evidence, maximum block thickness ( $z$ ) was taken as 10 m for block slides and 3 m for debris flows. For the materials occurring on low-angle dip slopes, friction angle was assigned a residual value of 10° (characteristic of marls) and for materials on scarp slopes the angle was assigned to be 25° (characteristic of sandstones). Cohesion was taken as 0.004 kN/m<sup>2</sup>, an average based on laboratory test results for the marl and mudstones on dip slopes, and 5 kN/m<sup>2</sup> for sandstones on scarp slopes (Bandis *et al.*, 1996). Eyewitness accounts indicate that a state of steady seepage occurred at surface level for some time after failure, so the ‘ratio of water table depth to failure surface’ or  $m$  can be

assumed to equal 1.0. ‘Map algebra’ was then used to calculate the infinite slope equation to generate a factor of safety ( $f$ ) map, using the following ‘map algebra’ expression, as constructed in ER Mapper’s formula editor:

```

if i1>0 and i1<20 or i1>240 and i1<360 then
(0.004 + (24 - 1*10)*(10*(cos(i1*(pi/180))
*cos(i1*(pi/180)))*tan(10*(pi/180))))/
(24*10*sin(i2*(pi/180)))
*cos(i2*(pi/180))) else
if i1>20 and i1<240 then
(5 + (24 - 1*10)*(3*(cos(i2*(pi/180))
*cos(i2*(pi/180)))*tan(25*(pi/180))))/
(24*10*sin(i2*(pi/180)))
*cos(i2*(pi/180))) else null

```

where  $i1$  = pixel values in the slope aspect image,  $i2$  = pixel values in the slope angle image, and null = pixels representing areas not at risk to failure and therefore excluded from the processing algorithm (no value).

4. Clay content and leached zones: Image ratios were derived from Landsat TM data to create two indices revealing selected ground characteristics: (i) iron oxide content, from TM bands 3/1; and (ii) hydrated mineral (including clay) content, from TM bands 5/7. The distribution of iron-oxide-rich areas of soil is included in the analysis as an indirect indication of the presence of highly fractured zones (where iron is preferentially leached from the rocks and soils around) and therefore of potential instability. The tasseled cap transform was used to produce a soil wetness index. This transform, used to derive indices such as ‘Brightness’, ‘Greenness’, and ‘Wetness’ from remotely sensed images, was developed by Crist and Cicone (1984a, 1984b). These three image-derived indices provide information about leaching and fracturing of the ground (iron oxide), the water-retentive properties of the soils (hydrated mineral and clay content) and soil moisture (wetness), which could then be used as evidence in the GIS geohazard assessment of conditions leading to instability.

**Table 21.13** Pairwise comparison matrix for factors influencing block-slide hazard

Factors	Slope	Aspect	FS	Rddist	Drdist	Wet	Fe	Clay
Slope	1	1	1	1/3	1/5	1/2	1	1/4
Aspect	1	1	1	1/3	1/9	1	1/3	1/5
FS	1	1	1	1/3	1/7	1/3	1/3	1/8
Rddist	3	3	3	1	1	3	3	1
Drdist	5	9	7	1	1	3	3	3
Wet	2	1	3	1/3	1/3	1	1/2	1/6
Fe	1	3	3	1/3	1/3	2	1	1/3
Clay	4	5	8	1	1/3	6	3	1
Factor weights	0.052	0.045	0.039	0.187	0.292	0.070	0.089	0.226

<sup>1</sup>Slope = gradient; FS = factor of safety; Rddist = distance from roads; Drdist = distance from drainage; Wet = wetness index; Fe = iron oxide index; and Clay = hydrated mineral index.

### 21.3.3.3 Multi-criteria evaluation

The model used here is based on the analytical hierarchy process described in Section 18.5.5, where there are both factors and constraints input to the model. The factors have been prepared using a variety of fuzzy membership functions and their significance was evaluated using a pairwise comparison matrix, as shown in Table 21.13. The factor weighting coefficients are then calculated using the method described in Section 18.4.3. For the purpose of illustration, we shall deal only with block-slides here.

The parameters controlling the fuzzy membership thresholds were selected on the basis of field

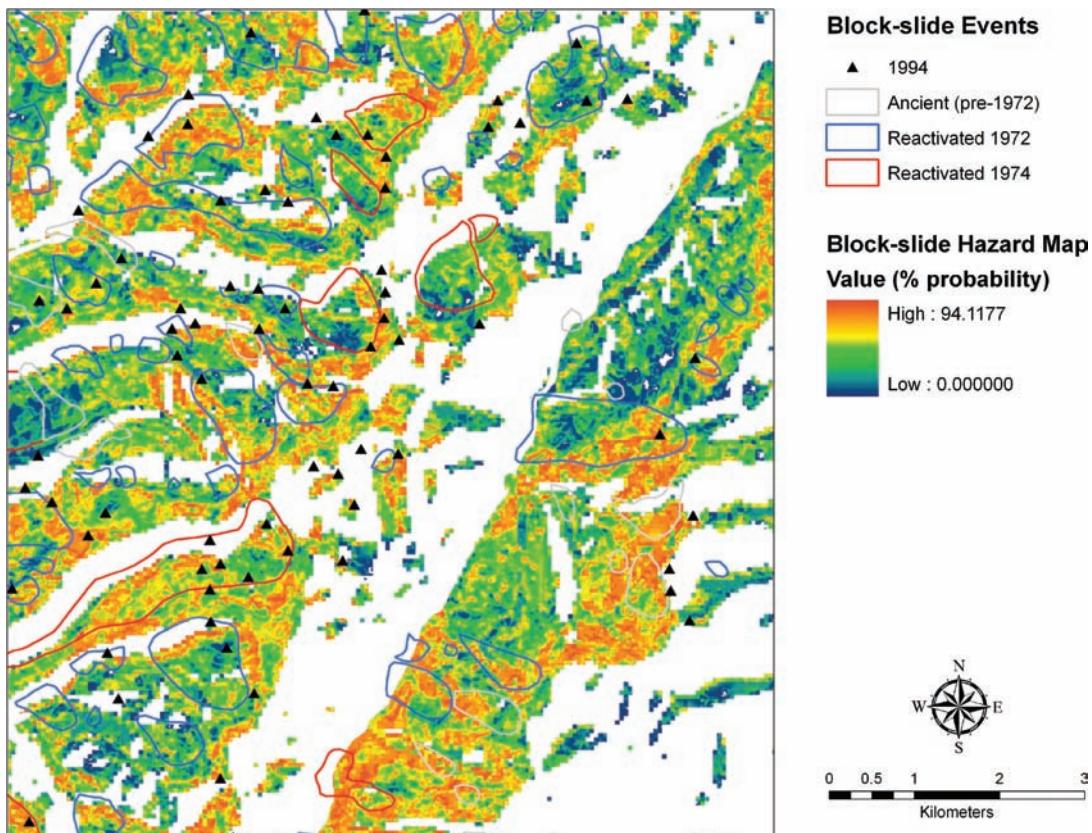
observation and published work, as described previously. These control points and the function types and forms used here are summarized in Tables 21.13 and 21.14.

### 21.3.3.4 Hazard and risk maps

The probability of occurrence of both spatial and temporal events needs to be determined with respect to the mass movement hazard. Varnes (1984) defines a hazard as being *the probability of occurrence of a potentially damaging phenomenon within a given time and in a given area*, and so we use the

**Table 21.14** Fuzzy membership functions used to prepare input criteria for block-slide hazard

Factors	Function type	Function shape	Control Point Values
Slope gradient	Sigmoidal	Monotonic decreasing	0, 4, 15, 20
Slope aspect	Sigmoidal	Symmetric	20, 40, 220, 260
Factor of safety	Linear	Monotonic decreasing	0, 1
Distance from roads	Sigmoidal	Monotonic decreasing	200, 1000
Distance from drainage	Sigmoidal	Monotonic decreasing	150, 300
Wetness	Sigmoidal	Monotonic increasing	80, 180
Iron-oxides	Sigmoidal	Monotonic increasing	50, 150
Hydrated minerals	Sigmoidal	Monotonic increasing	70, 200



**Figure 21.29** Block-slide hazard map, representing probability of occurrence, as a worst case scenario

relationship between hazard, risk and vulnerability described in Section 17.4.

As stated in Section 17.2, decisions concerning the hazard being considered can be computed within GIS by employing rules based on logic. Where data values have been measured directly, ‘hard’ decision rules can be formulated. This is difficult to achieve in reality and generally ‘soft’ decisions have to be established on the basis of experience, prior knowledge and judgement; in other words, ‘belief’ in the possible outcomes.

Uncertainty in this case describes both the natural variability of the data and the lack of evidence about the significance of the data. This can be extended to consider whether a slope could become unstable as a result of an adverse combination of parameters.

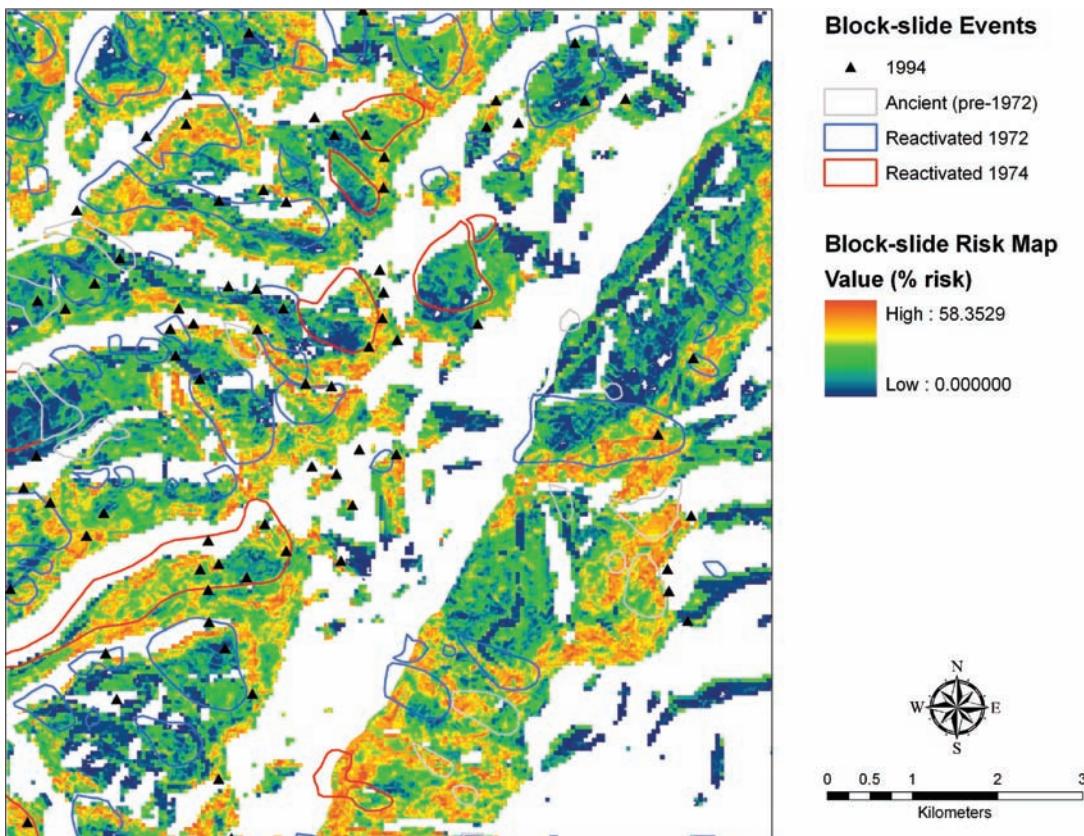
The hazard map for block slides is shown in. Much of the area has probability values exceeding 0.05, and the NW-facing slopes generally show

values greater than 0.10. Comparison between distributions of block slides in 1972, 1974 and 1994 (shown in Figure 21.29) reveals coincidence with many areas of highest hazard.

A map of relative vulnerability can be deduced through reclassification of the attributes to derive generalized land-use classes, and assigning them values between 0 and 1 as measures of their relative cost value (with 1 representing the highest vulnerability). The hazard maps for block slides and debris flows can then be multiplied by the vulnerability map to produce landslide risk maps, as shown in Figure 21.30.

### 21.3.4 Summary

The study of slope stability and geohazard assessment has attracted a great deal of attention as concern



**Figure 21.30** Block-slide risk map, representing percentage risk (under the same conditions as above) according to the relative value of land and property in the area

has grown for the safety of urban development encroaching on upland areas. Landslides in this area generally involve translational sliding and so the application of the infinite slope method proved an acceptable model for this study. The results show that planar failures are more likely to occur on the NW-facing dip slopes, but that if the soil/rock interface is taken as a potential discontinuity, planar failures may also occur on scarp slopes.

Comparison of the hazard maps generated by GIS with the distribution of known landslide events has revealed the general applicability of the methodology. It is acknowledged, however, that the database used in this work is incomplete and contains errors, and that work in this area has continued (Luino, 1999; Godio and Bottino, 2001; Guzzetti, 2000; Canuti *et al.*, 2004; Meisina *et al.*, 2008). Furthermore the planar, infinite slope model is known to be a simplification of the actual failure mechanisms operating but the geohazard map computed in this manner seems to

reflect reasonably the observed occurrences of landslides in the Langhe region.

Image information relating to landslides is complex and contains two important components: texture and spectral detail. Remote sensing has been widely and successfully used to detect landslides in the past (Murphy and Vita-Finzi, 1991; Rengers, Soeters and Van Weston, 1992; Murphy and Bulmer, 1994). It also provides a very convenient source of time-dependant information.

Geomorphological studies indicate that the Langhe region is still dynamic in terms of post-Alpine, post-glacial crustal uplift and that the mass movements are an ongoing natural, slope-dynamic consequence of this uplift. Recent research also points to a close link between slope instability in the Alps and periodicity in the Holocene climate (Canuti *et al.*, 2004). This implies that landsliding in the region is a long-standing phenomenon and is likely to remain so. Such situations are not uncommon, so the

continuing development and exploitation of new technologies to help understand and mitigate the effects of such geohazards are vitally important.

The hypothesis that incorporation of digital information within geohazard assessment utilizing GIS can significantly improve risk management in areas such as the Langhe has been considered for:

- compilation of thematic information from remote sensing, geomorphology (elevation and its derivatives) and land usage;
- stability analysis of selected slope profiles and of the whole study area;
- multi-criteria hazard assessment (using probability and decision support tools) to compute geo-hazard maps.

It can be concluded that for image-based studies utilizing satellite data, the most significant information is morphological since it is these features that are detectable using the sensor. There are certain important features that are needed for the correct identification of mass movements, for example arcuate scarp, tension cracks and hummocky displaced ground. Detection of such features in imagery is helped by prior knowledge of the likely mechanisms and the prevailing state of activity.

GIS provides a flexible and effective tool for slope stability assessment and the production of thematic maps. The multi-criteria approach provides a practical means for aggregating significant attributes (factors) influencing slope instability and also provides a flexible means of combining individual factors reflecting their relative influence on the system controlling the outcome.

## 21.4 Land surface change detection in a desert area in Algeria using multi-temporal ERS SAR coherence images

As indicated in Chapter 10, multi-temporal SAR interferometric coherence imagery is a useful information source for the detection of random changes of the land surface. In this case study, three coherence images derived from three ERS-1 SAR images of an arid area of the Sahara Desert in Algeria revealed some interesting phenomena, including distribution

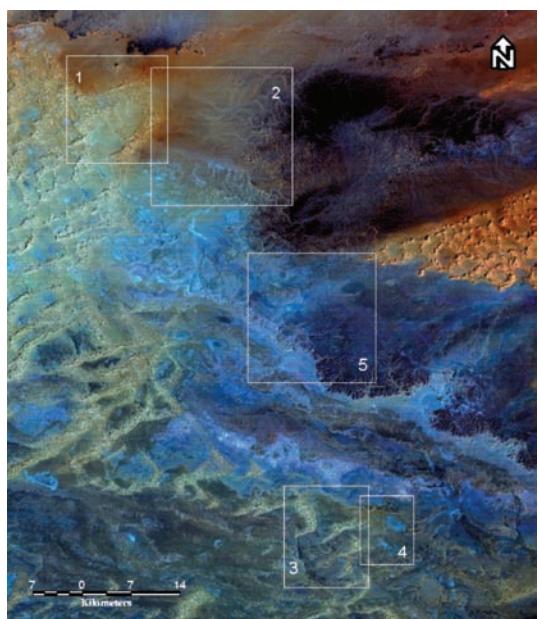
of mobile sand, erosion along river channels, variation of ephemeral lakes, and seismic survey lines.

### 21.4.1 The study area

The area chosen for study is in eastern Algeria near the border with Libya in North Africa,  $100 \times 100 \text{ km}^2$ , at approximately  $27\text{--}28^\circ\text{N}$  and  $8\text{--}9^\circ\text{E}$ . The Atlas Mountains separate the warm and temperate region along the coast of the Mediterranean from the vast hot arid desert: the Sahara. With very low humidity levels from 5 to 25%, rainfall is rare, solar radiation is intensive and the diurnal variation of temperature is large in the region (Ahrens, 1994).

The very low precipitation and excessive evaporation make the desert hyper-dry, barren and almost completely devoid of surface vegetative cover. This absence of a binding agent allows the loose sand or topsoil to migrate according to the prevailing wind patterns. It has been observed that the desert in this region is expanding northwards, with the vegetation of marginal lands being stripped for firewood or animal fodder, further exposing fragile soils to erosion.

As shown in a colour composite of a Landsat TM image (Figure 21.31), the main geographic features of the study area are large expanses of flat bare rock or gravel plains broken up by escarpments, gully



**Figure 21.31** Study area: Landsat TM colour composite of bands 4, 2 and 1 in RGB (10 February 1987)

networks and ephemeral drainage channels, some of which flow into lakes or depressions. Large parts of the region are covered with seas of sand, with linear, barchanoid and star dune types present, as well as thin sand sheets.

#### 21.4.2 Coherence image processing and evaluation

Three scenes of ERS-1 SAR raw data of the study area acquired on 8 September 1992, 13 October 1992 and 28 September 1993 were processed by an SAR processor to produce single look complex (SLC) images. We name the three scenes in time sequence as *Alg1*, *Alg2* and *Alg3*. Among the three SLC images, *Alg2* was used as the master scene while *Alg1* and *Alg3* were used as slave scenes to be co-registered to the master. Three coherence images with 35, 350 and 385 days of temporal separation were thus produced using formula (10.23) and named as *Coh12*, *Coh23* and *Coh13*.

There are several decorrelation factors that cause the loss of coherence in multi-temporal coherence imagery (Zebker and Villasenor, 1992; Gens and van Genderen, 1996). Besides the temporal change of the land surface, which is the target of the study, the major factors reducing coherence level are the baseline distance and the local slope.

Decorrelation caused by baseline separation is an inherent factor of the multi-pass, multi-temporal interferometric SAR system. The component of baseline perpendicular to radar look direction ( $B_{\perp}$ ) represents the difference in view angles for the same ground target between the two observations. The

phase of a radar return signal is decided by the vector summation of all the scatterers within a ground resolution cell. If  $B_{\perp}$  is significant, the radar beam will illuminate the same ground target at considerably different angle and the collective effects of the relevant scatterers will result in a certain degree of random variation of phase. Thus the coherence decreases with the increase of  $B_{\perp}$  as characterized in the formula (10.24) in Chapter 10 and thus a short  $B_{\perp}$  is generally preferred for coherence-based random change detection.

We can calculate the theoretical coherence values of the three coherence images from formula (10.24) using the nominal parameters of the ERS-1 SAR system. These data together with the actual average coherence values of the whole scene, a high coherence flat area and a gully-dissected area are shown in Table 21.15. The theoretical coherence value declines steadily with the increase of  $B_{\perp}$ . The average coherence for the whole scene is much lower than the theoretical value for all the three coherence images because of very low coherence resulting from temporal decorrelation in the large areas covered by mobile sand. It is interesting to notice that the average coherence over the gully-dissected area for *Coh23* is higher than that for the full scene and it is significantly higher than those for *Coh12* and *Coh13* which are lower than their corresponding full-scene averages.

For further analysis, the ratios between the actual and the theoretical coherence values were calculated (Table 21.15). The ratio data give an evaluation of the relationships among  $B_{\perp}$ , local slope, temporal changes and coherence level. For the full scene, the  $\rho_{\text{actual}}/\rho_{\text{theory}}$  ratio declines with the increase of

**Table 21.15** Coherence data of the three coherence images of the study area

		<i>Coh12</i>		<i>Coh23</i>		<i>Coh13</i>	
Time Separation (days)		35	$\frac{\rho_{\text{actual}}}{\rho_{\text{theory}}}$	350	$\frac{\rho_{\text{actual}}}{\rho_{\text{theory}}}$	385	$\frac{\rho_{\text{actual}}}{\rho_{\text{theory}}}$
$B_{\perp}$ (metre)		263		105		368	
Theoretical $\rho_{\text{theory}}$		0.7958		0.9185		0.7143	
Actual $\rho_{\text{actual}}$	Full scene	0.5142	0.6461	0.5569	0.6063	0.4279	0.5990
	High coherence area	0.6677	0.8390	0.7939	0.8643	0.5460	0.7644
	Gully area	0.3900	0.4901	0.6104	0.6646	0.3496	0.4894

**Table 21.16** Coherence scenarios and logical interpretations

Scenario	Coherence level		Interpretation
	<i>Coh12</i> (35 d)	<i>Coh23</i> (350 d)	
	8 Sept. to 13 Oct. 1992	13 Oct. 1992 to 28 Sept. 1993	
1	High	High	Stable, no change.
2	High	Low	Stable, then substantial change after 13 Oct. 1992
3	Low	High	Sudden change within the first 35 d then stable in the following 350 d
4	Low	Low	Continual substantial change over whole period. <i>Or</i> sudden change in 35 d followed by substantial change in 350 d
5	Medium	High	Slight change until 13 Oct. 1992 and then stable
6	Medium	Low	Slow and progressive change over whole period

temporal separation because the random changes of land surface accumulate with time. In the gully area, however, the value of  $\rho_{\text{actual}}/\rho_{\text{theory}}$  increases significantly between *Coh12* and *Coh23* when  $B_{\perp}$  decreases from 263 to 105 m. This is due to the spatial decorrelation effect on directly radar-facing slopes (Lee and Liu, 1999). For the same reason, the  $\rho_{\text{actual}}/\rho_{\text{theory}}$  ratio decreases gently with the increase of  $B_{\perp}$  in the flat stable area.

From Table 21.15, it is obvious that *Coh13* has the poorest quality of coherence because of the largest  $B_{\perp}$  among the three coherence images, and it covers the repeated temporal range of *Coh12* and *Coh23*. Therefore, *Coh13* is only used when necessary in the following interpretation for change detection.

### 21.4.3 Image visualization and interpretation for change detection

#### 21.4.3.1 Principles of interpretation

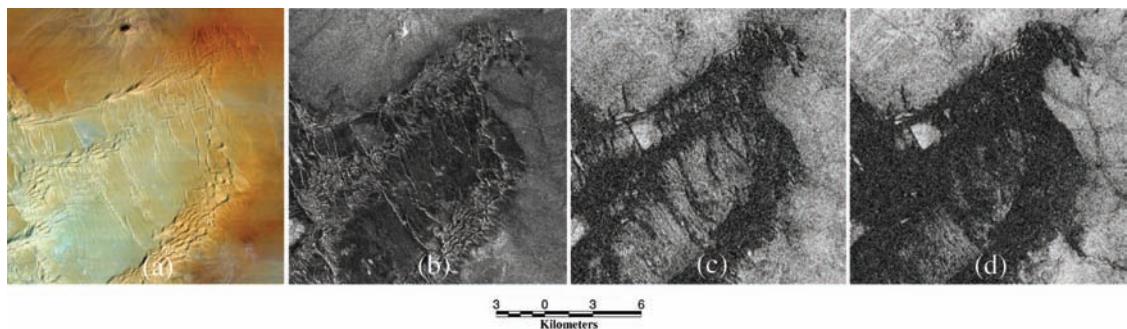
The study area has a very stable environment. The possible factors causing random changes of land surface are sand movement, erosion and deposition caused by wind or occasional flash flooding and limited human activities mainly relating to oil exploration. These changes will cause the decrease and loss of coherence and form dark features remarkably obvious against the high-coherence background of a stable barren land surface. With three images taken with 35 and 350 and 385 days

of temporal separation, simple logical analysis is effective for interpreting the nature of the changes. Typically, there are six possible scenarios between *Coh12* and *Coh23*, as shown in Table 21.16.

#### 21.4.3.2 Sand movement (Boxes 1 and 3 in Figure 21.31)

Several types of sand dunes are present in the area, including transverse barchan and linear types, and star dune networks. These are generally evident on TM imagery (Figure 21.32), which shows the morphology and structure of individual dune features. However, to define the boundaries of a dune or dune field and to identify thin sheets of mobile sand are not always possible using TM or other types of optical imagery, particularly when the spectral properties of sand are very similar to the solid basement, as shown in Figure 21.32a. SAR amplitude imagery is even less adequate for the task, as shown in Figure 21.32b, because the tone variation of the image is relevant to surface roughness rather than spectral or dynamic properties.

In contrast, based on quite different principles, coherence imagery is very effective for dune boundary delineation and mobile sand sheet identification, thus enabling a critical assessment of dune movement and sand encroachment. The loose sand grains on dune surfaces or thin sand sheets on a solid basement plain are subject to continuous movement under the wind even though the dune is



**Figure 21.32** Dune boundary definition and mobile sand detection: (a) colour composite of Landsat TM band 421 in RGB (box 1 in Figure 21.31); (b) multi-look amplitude image of *Alg1* scene; (c) *Coh12*, the coherence image of 35 days of separation; and (d) *Coh23*, the coherence image of 350 days of separation

static as a whole. The sand movement causes random changes in the micro-geometry of scatterers on the sand-covered land surfaces and thus results in a loss of coherence over a very short period as characterized by scenario 4 in Table 21.16. The very dark decoherence features of mobile sand over a bright high-coherence background are not only direct evidence of sand mobility, but also effectively delineate the outlines of active dunes and optically indiscernible thin sand sheets. These data are not easily obtainable over a large region using other Earth observation techniques.

As illustrated in Figure 21.32c, complex boundaries of three chains of dunes (barchan and linear types) in the region are sharply defined in the *Coh12* image as decoherence patches over a high-coherence background. The boundaries are distinctive and definite. With 350-day temporal separation, *Coh23* (Figure 21.32d) reveals a thin sheet of mobile sand spreading into the inter-dune areas making the whole dune field a nearly continuous decoherence patch. The central part of the dune field is typically characterized by scenario 6 in Table 21.16 as medium coherence in *Coh12* and low coherence in *Coh23* indicating continuous transport of the sand sheet as it is swept over the barren land surface.

The dune positions are defined effectively in the coherence images of 35, 350 and 385 days of temporal separation. A colour composite of the three coherence images may reveal possible dune migration, which occurred during the 385-day period. For a colour composite of *Coh12* in red, *Coh23* in green and *Coh13* in blue, a quickly migrating

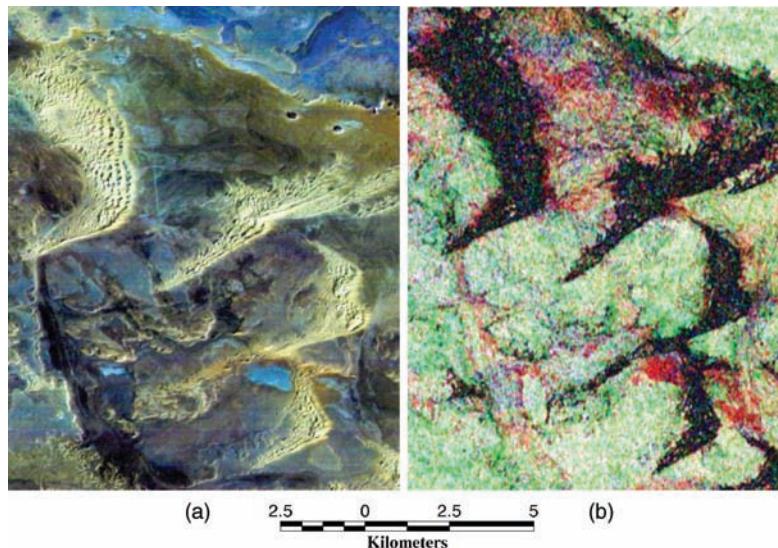
barchan dune would be presented as a dark decoherence feature with a narrow trailing edge in red and a windward edge in green (Liu *et al.*, 1997a). As illustrated in Figure 21.33b, this diagnostic pattern is not evident, a discovery not unexpected for the following reasons:

- The large formations approximately 1 km wide are static as a whole. These large formations consist of small barchanoid ridges 50 m wide, which themselves are likely to be migrating features, but the migration cannot be detected in the largely decoherent background of the large sand formations.
- Even dunes migrating rapidly at 20 m per year would not produce a substantial signal on the images of 35- and 350-day intervals, at a coarse pixel resolution of around 30 m.

In order to make a serious attempt to identify dune migration, coherence imagery with a much longer temporal separation is required.

#### 21.4.3.3 Ephemeral lakes and water bodies (Box 4 in Figure 21.31)

The RGB colour composite of coherence images is an effective aid for the logical analysis of various events of land surface changes. The area defined by the box in the coherence colour composite of Figure 21.34a presents an obvious red patch. It appears to correspond strongly with a bright cyan feature on the TM 421 colour composite (Figure 21.34d), which is defined as a shallow ephemeral lake in a reference map of the area



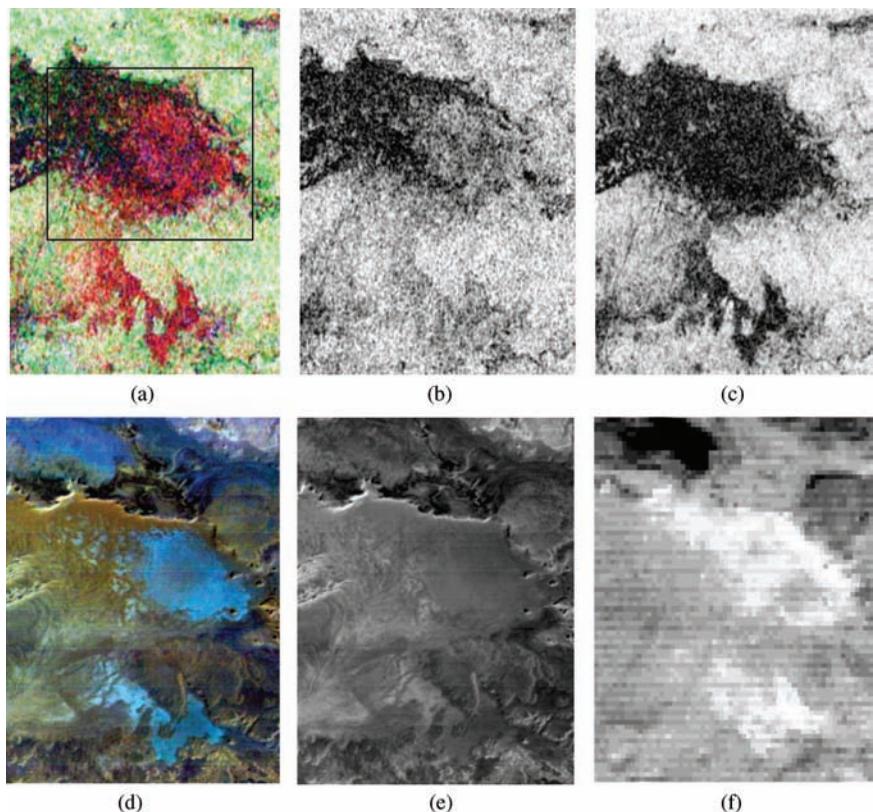
**Figure 21.33** Dune migration: (a) colour composite of Landsat TM bands 421 in RGB (box 3 in Figure 21.31); (b) colour composite of *Coh12*, *Coh23* and *Coh13* in RGB; a migrating barchan dune would be presented as a dark decoherence feature with a narrow trailing edge in red and windward edge in green. Note that this diagnostic feature does not appear in the image

(DMAAC, 1981). The analysis of TM multispectral information indicates that the lake was nearly dry when the TM image was taken on 10 February 1987 (there was no precipitation in January and February 1987 according to the data from the GPCC website). As shown in Figure 21.34e and f, the lake patch is not particularly dark in the NIR band *TM4* and very bright in the thermal band *TM6*. This characteristic is contradictory to the typical water spectral signature: strong absorption in *TM4*. The area in fact presents an unusual spectral property: high albedo and high thermal emission. In general cases, high-albedo objects would have low thermal emission (Liu, Moore and Haigh, 1997b). The exceptional cases may occur for crystallized transparent material with strong internal scattering such as snow, gypsum and salt. It is reasonable to presume that this dried saline lake basin is covered with salt deposits.

The red pixels in Figure 21.34a are those coherent in *Coh12* (Figure 21.34b), but not in *Coh23* (Figure 21.34c), logically implying a slow-changing environment that appears relatively stable in the short term (35 days) but the accumulated progressive change is substantial over a much longer period (350 days). It is therefore suggested that the lake

basin was dry during the initial 35 days with a relatively stable surface. This condition allows medium to high coherence in *Coh12*. Then, in the following 350 days, the lake possibly experienced recharges of floodwater, temperature variation over a considerable range and repeated salt mineral crystallization due to the water-level change. Any of these processes can produce random changes significant enough to result in decoherence in *Coh23*. This explanation is supported by monthly average precipitation data (Rudolf *et al.*, 1994; GPCC 1992–1993) of the area during the period as shown in Figure 21.35. There were 6–10 mm of precipitation in winter 1992 and 5–6 mm in autumn 1993, which is adequate to cause seasonal recharge to the lake.

Numerous similar patchy features can be identified in this region using the same logic and methodology, which correspond well with ephemeral lakes in the *TM421* RGB colour composite (Figure 21.34d). Obviously, a confident identification of these desert lakes cannot be achieved without the TM colour composite. The extra contribution of the SAR multi-temporal coherence image is the detection of the dynamic activities of these lakes.

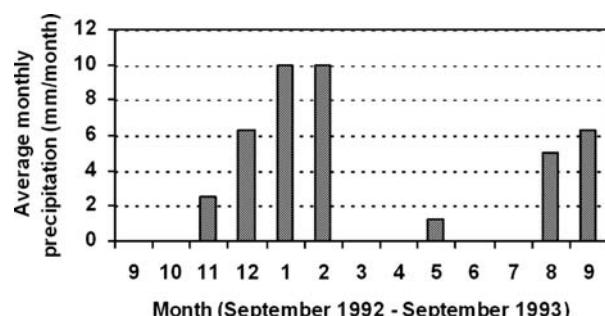


**Figure 21.34** The spectral and coherence properties of an ephemeral lake. (a) Colour composite of *Coh12*, *Coh23* and *Coh13* in RGB. The rectangular box indicates an area with progressive decreasing of coherence which is better shown by comparison between (b) the *Coh12* image and (c) the *Coh23* image where the patch is becoming darker. (d) Colour composite of Landsat TM bands 421 in RGB (box 4 in Figure 21.31) indicates that the feature is in cyan colour and likely a water body; however, lack of water absorption in (e) the TM band 4 image and strong thermal emission in (f) the TM band 6 image indicates there was no water when the image was taken

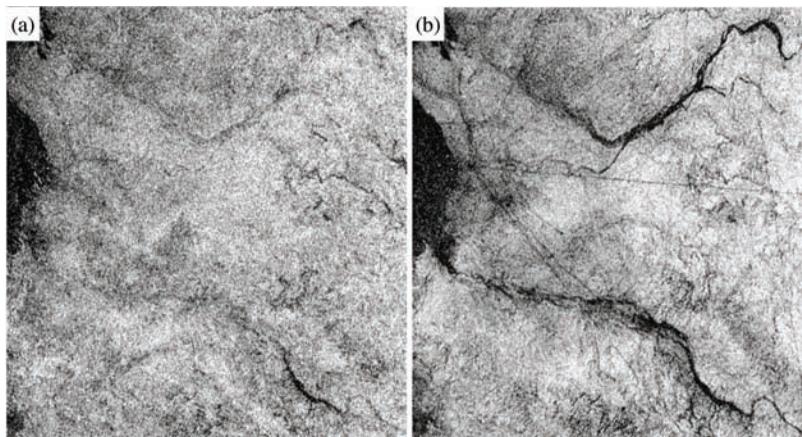
#### 21.4.3.4 Drainage pattern and erosion (Box 2 in Figure 21.31)

Though the dominant agent of erosion in the Sahara Desert is the prevailing wind, occasional

and isolated intense rainstorms can cause local flooding and rapid fluvial erosion/deposition. Multi-temporal coherence imagery can provide direct evidence of this process. As there is an acute



**Figure 21.35** Average monthly precipitation from September 1992 to September 1993 for the study area in the Sahara ( $E7^{\circ}-9^{\circ}$ ,  $N27^{\circ}-28^{\circ}$ ). Data compiled from Global Precipitation Climatology Centre: <http://www.dwd.de/research/gpcc>



**Figure 21.36** Drainage channels: two major rivers in the region are not visible in (a) for the 35-day separation coherence image *Coh 12* but clearly shown as decoherence features in (b) for the 350-day separation coherence image *Coh 23*

lack of information on the spatial location and temporal frequency of such erosion/deposition events, coherence imagery represents a valuable potential source of such data.

As shown in Figure 21.36a, the coherence image *Coh12* illustrates an area with high coherence over the initial 35-day period. There are no obvious water channel features except for a small section of channel approximately 30 m wide in the bottom-right corner of the image with low coherence. The subsequent *Coh23* image (Figure 21.36b), on the other hand, exhibits two separate major channels as obvious decoherence features in a bright background of high coherence. These features are very eminent from the east and gradually become less pronounced further downstream towards the west. This characteristic reflects localized flooding from isolated storms, coupled with high transmission losses and evaporation causing surface flow to diminish downstream. As shown in Figure 21.35, there was no precipitation during September and October 1992. We can therefore assume that the channels were dry and stable in the initial 35 days. In the subsequent 350 days, there were 6–10 mm of precipitation in winter 1992 and 5–6 mm in autumn 1993. These rainfall events could have caused seasonal flash floods in the rivers and resulted in active erosion/deposition resulting in loss of coherence in the *Coh23* image.

#### 21.4.3.5 Geophysical survey lines (Box 5 in Figure 21.31)

We have alluded to this unexpected finding in Chapter 10. As presented in Figure 10.8, the coherence image *Coh23* observed between 13 October 1992 and 28 September 1993 reveals a mesh of straight lines which are not present in the relevant SAR multi-look images. These lines are clearly visible over this approximately  $40 \times 40 \text{ km}^2$  area, with single lines up to 40 km long at a spacing of 2 to 3 km. Obviously, these are the results of anthropogenic disturbance over the periods between the repeated SAR image acquisitions. It is known that Sonatrach, the national oil company of Algeria, operated in the area during the period; the coherence image *Coh23* exposed considerable details of seismic survey.

A seismic survey line will result in decorrelation between the SAR images taken before and after the survey as long as the swath is equivalent to or wider than the SAR image resolution cell and the random disturbance caused by engineering work is greater than half a wavelength of the radar beam (2.83 cm) in its slant range direction. The features are therefore detectable using the coherence image in such a largely stable environment. However, the disturbance to the ground, after land surface recovering, is not great enough to alter significantly the average intensity of return SAR signals corresponding to each pixel of an SAR

multi-look image, thus the features are not visible in Figure 10.8a.

This case demonstrates a unique function of coherence imagery as a tool for monitoring the environmental impact of human activities.

#### 21.4.4 Summary

The primary value of coherence imagery lies in its ability to record efficiently very subtle random changes on the land surface in an otherwise stable environment. If the average random change of dominant scatterers within a resolution cell exceeds one-half of the radar wavelength in the slant range direction, it will cause total decorrelation. Changes at this small scale are usually not detectable on conventional optical imagery. The change detection technique based on multi-temporal SAR coherence imagery is fundamentally different from the SAR interferogram-based measurement technique. Differential SAR interferometry is capable of measuring centimetre-level land surface deformation (a consistent block movement) but not workable with the random changes in the same scale range. On the other hand, coherence imagery can detect the centimetre-level random changes but cannot provide quantitative measurements.

In an arid environment such as the Sahara discussed here, the predominantly bare desert surface forms an extremely stable landscape which retains high coherence over very long periods (several years). The contrast between this bright background and the dark decorrelation signatures of any random changes enables the detection and delineation of unstable features. It is this property which facilitates the spatial and temporal mapping of surface processes with a confidence unrivalled by other Earth observation techniques, and over areas too large or inaccessible for effective field surveys. This case study demonstrates clearly the potential of SAR coherence imagery to detect and interpret changes in a desert environment. For example, persistent decorrelation over short time intervals is direct evidence of sand mobility.

The lack of precipitation data in remote desert regions often hampers attempts to research the contribution of catastrophic fluvial erosion to arid landscapes. A sequence of short time-scale (month-

ly), frequent coherence images could provide critical objective information on the temporal and spatial distribution of localized sporadic flood events. Coherence imagery also provides an effective way to detect human-induced disturbances over various time intervals.

A multi-temporal SAR coherence image presents an objective record of irregular land surface changes between two SAR image acquisitions as decoherence features. Such low-coherence phenomena can easily be distinguishable only when they are in sharp contrast to a high-coherence background. The technique is most effective for detecting changes in a largely stable environment, such as desert, but needs more sophisticated analysis in an unstable environment with many other decorrelation factors.

## Questions

### Section 21.3

- 21.1 Is the WLC method the most appropriate to use in a case like this?
- 21.2 What other datasets could be considered?
- 21.3 Why do the extents of the older landslides not match the areas predicted in the hazard assessment and risk maps?
- 21.4 What can we say about the temporal constraints on a hazard prediction of this kind?

## References

### Section 21.1

- Cihlar, J., St-Laurent, L. and Dyer, J.A. (1991) Relation between the normalized difference vegetation index and ecological variables. *Remote Sensing of Environment*, **35** (2), 257–277.
- Gausman, H.W. (1974) Leaf reflectance of near-infrared. *Photogrammetric Engineering and Remote Sensing*, **10**, 183–191.
- Lambin, E.F. and Ehrlich, D.E. (1997) Land cover changes in Sub-Saharan Africa (1982–1991): application of a change index based on remotely sensed surface temperature and vegetation indices at a continental scale. *Remote Sensing of Environment*, **61** (2), 181–200.

- Lee, E.M. and Jones, D.K.C. (2004) *Landslide Risk Assessment*, Thomas Telford, London.
- Li, B., Tao, S. and Dawson, R.W. (2002) Relations between AVHRR NDVI and ecoclimatic parameters in China. *International Journal of Remote Sensing*, **23**, 989–999.
- Lilesand, T.M. and Kiefer, R.W. (2000) *Remote Sensing and Image Processing*, 4th edn, John Wiley & Sons, Inc., New York.
- Liu, J.G. and Meng, M. (2005) Destruction of vegetation in the catchments of Nujiang river, 'Three Parallel Rivers' region, China. Proceedings of the 2005 IEEE International Geoscience and Remote Sensing Symposium (IGARSS 2005), 25–29 July 2005, Seoul, Korea.
- Mantovani, A.C.D. and Setzer, A.W. (1997) Deforestation detection in the Amazon with an AVHRR-based system. *International Journal of Remote Sensing*, **18**, 273–286.
- Meentemeyer, R.K. and Moody, A. (2000) Automated mapping of conformity between topographic and geological surfaces. *Computers & Geosciences*, **26** (7), 815–829.
- Natural site datasheet from WCMC:  
<http://www.unep-wcmc.org/sites/wh/pdf/THREE%20RIVERS%20YUNNAN.pdf>  
 (accessed 20 February 2009).
- Wang, J., Price, K.P. and Rich, R.M. (2001) Spatial patterns of NDVI in response to precipitation and temperature in the central Great Plains. *International Journal of Remote Sensing*, **22**, 3827–3844.
- Wen, B.P., Wang, S.J., Wang, E.Z. and Zhang, J.M. (2004) Characteristics of rapid giant landslides in China. *Landslides*, **1**, 247–261.
- World Heritage nomination – IUCN technical evaluation, Three Parallel Rivers of Yunnan Protected Areas (China), ID No. 1083 (2003) [http://whc.unesco.org/archive/advisory\\_body\\_evaluation/1083.pdf](http://whc.unesco.org/archive/advisory_body_evaluation/1083.pdf). (accessed 20 February 2009).
- World Heritage 27 COM, WHC-03/27.COM/24, Paris (10 December 2003), pp. 91–92.
- Fourniadis, I.G., Liu, J.G. and Mason, P.J. (2007a) Regional assessment of landslide impact in the Three Gorges area, China, using ASTER data: Wushan-Zigui. *Landslides*, **4**, 267–278.
- Fourniadis, I.G., Liu, J.G. and Mason, P.J. (2007b) Landslide hazard assessment in the Three Gorges area, China, using ASTER imagery: Wushan-Badong. *Geomorphology*, **84** (1–2), 126–144.
- Guzzetti, F., Reichenbach, P., Cardinali, M., Ardizzone, F. and Galli, M. (2003) The impact of landslides in the Umbria region, central Italy. *Natural Hazards and Earth System Sciences*, **3** (5), 469–486.
- Hartlen, J. and Viberg, L. (1988) General report: evaluation of landslide hazard. Proceedings of the Fifth International Symposium on Landslides, Lausanne, pp. 1037–1057.
- Huang, R. and Li, Y. (1992) Logical model of slope stability prediction in the Three Gorges Reservoir area, China. Proceedings of the Sixth International Symposium on Landslides-Glissemens de terrain, Christchurch, pp. 977–981.
- Huang, J., Xie, S. and Kuang, M. (2001) Geomorphic evolution of the Yangtze Gorges and the time of their formation. *Geomorphology*, **41**, 125–135.
- Hutchinson, J.N. (1995) Keynote paper: Landslide hazard assessment, in *Landslides: Sixth International Symposium on Landslides* (ed. D.H. Bell), Balkema, Rotterdam, pp. 1805–1841.
- Li, J., Xie, S. and Kuang, M. (2001) Geomorphic evolution of the Yangtze Gorges and the time of their formation. *Geomorphology*, **41**, 125–135.
- Liu, J.G., Mason, P.J., Clerici, N. et al. (2004) Landslide hazard assessment in the Three Gorges Area of the Yangtze River using ASTER imagery: Zigui-Badong. *Geomorphology*, **61** (1–2), 171–187.
- Mantovani, F., Soeters, R. and Van Westen, C.J. (1996) Remote sensing techniques for landslide studies and hazard zonation in Europe. *Geomorphology*, **15**, 213–225.
- Mason, P.J. and Rosenbaum, M.S. (2002) Predicting future landslides in a residential area on the basis of geohazard mapping: the Langhe Hills in Piemonte, NW Italy. *Quarterly Journal of Engineering Geology and Hydrology*, **35**, 317–326.
- Saha, A.K., Gupta, R.P. and Arora, M.K. (2002) GIS-based landslide hazard zonation in the Bagirathi (Ganga) Valley, Himalayas. *International Journal of Remote Sensing*, **23**, 357–369.
- Varnes, D. (1984) Landslide hazard zonation: a review of principles and practice. Commission on Landslides of the International Association of Engineering Geology, United Nations Educational Social and Cultural Organisation, Natural Hazards, No. 3.

## Section 21.2

- Chen, Q., Hu, H., Sun, Y. and Tan, C. (1995) Assessment of regional crustal stability and its application to engineering geology in China. *Episodes*, **18**, 69–72.
- Chen, S. (1986) Atlas of Geo-Science Analyses of Land-sat Imagery in China, National Remote Sensing Centre, Chinese Academy of Science Science Press, Beijing.
- China Yangtze Three Gorges Project Development Corporation (1999) The Three Gorges Project, 1999 report.

- Welch, R., Jordan, T., Lang, H. and Murakami, H. (1998) ASTER as a source for topographic data in the late 1990's. *IEEE Transactions on Geoscience and Remote Sensing*, **36**, 1282–1289.
- Wu, S., Hu, D., Chen, Q. *et al.* (1997) Assessment of the crustal stability in the Qingjiang river basin of the western Hubei Province and its peripheral area, China. Proceedings of the Thirtieth International Geological Congress, pp. 375–385.
- Wu, S., Shi, L., Wang, R. *et al.* (2001) Zonation of the landslide hazard in the forereservoir region of the Three Gorges Project on the Yangtze River. *Engineering Geology*, **59**, 51–58.
- Yamaguchi, Y., Fijisada, H., Tsu, H. *et al.* (2001) ASTER early image evaluation. *Advanced Space Research*, **28**, 69–76.
- Zhao, C. (1996) River capture and origin of the Yangtze River. *Journal of Changchun University of Earth Sciences*, **26**, 428–433.
- data. *Photogrammetric Engineering and Remote Sensing*, **50** (3), 343–352.
- Embleton, C. (ed) (1984) *Geomorphology of Europe*, Macmillan, London.
- Forlati, F., Lancellotta, R., Osella, A. *et al.* (1996) The role of swelling marl in planar slides in the Langhe region, in *Landslides*, vol. 2 (ed. K. Senneset), Balkema, Rotterdam, pp. 721–725.
- Godio, A. and Bottino, G. (2001) Electrical and electromagnetic investigation for landslide characterization. *Physics and Chemistry of the Earth, Part C: Solar, Terrestrial & Planetary Science*, **26** (9), 705–710.
- Govi, M. (1974) La frana di Somano (Langhe Cuneesi). *Studi Trentati di Scienze Naturale*, **51**, 153–165.
- Govi, M. and Sorzana, P.F. (1982) Frana di scivolamento nelle Langhe Cuneesi Febbraio-Marzo 1972, Febbraio 1974. *Bulletina della Associazione Mineraria Sub-alpina, Anno*, **XIX** (1–2), 231–263.
- Guzzetti, F. (2000) Landslide fatalities and the evaluation of landslide risk in Italy. *Engineering Geology*, **58** (2), 89–107.
- Luino, F. (1999) The flood and landslide event of November 4–6, 1994 in Piedmont Region (Northwestern Italy): causes and related effects in Tanaro Valley. *Physics and Chemistry of the Earth, Part A*, **24** (2), 123–129.
- Mason, P.J., Palladino, A.F. and Moore, J.McM. (1996) Evaluation of radar and panchromatic imagery for the study of flood and landslide events in Piemonte, Italy, in November 1994. Proceedings of the European School on Floods and Landslides: Integrated Risk, 19–26 May, Orvieto, Italy.
- Mason, P.J. and Rosenbaum, M.S. (2002) Geohazard mapping for predicting landslides: the Langhe Hills in Piemonte, NW Italy. *Quarterly Journal of Engineering Geology & Hydrology*, **35**, 317–326.
- Mason, P.J., Rosenbaum, M.S. and Moore, J.McM. (2000) Predicting future landslides with remotely sensed imagery, in *Landslides in Research, Theory and Practice* (eds E. Bromhead, N. Dixon and M. Ibsen), Proceedings of the 8th International Symposium on Landslides, Cardiff, 26–30 June, Thomas Telford, London, pp. 1029–1034.
- Meisina, C., Zucca, F., Notti, D. *et al.* (2008) Potential and limitation of PSInSAR technique for landslide studies in the Piemonte Region (Northern Italy). *Geophysical Research Abstracts*, **10**, EGU2008-A-09800.
- Murphy, W. and Bulmer, H.K. (1994) Evidence of prehistoric seismicity in the Wairarapa Valley, New Zealand, as indicated by remote sensing. Proceedings of the Tenth Thematic Conference on Geologic Remote Sensing, San Antonio, Texas, 9–12 May, I, pp. 341–351.

### Section 21.3

- Bandis, S.C., Delmonaco, G., Margottini, C. *et al.* (1996) Landslide phenomena during the extreme meteorological event of 4–6 November 1994 in Piemonte Region in N. Italy, in *Landslides*, vol. 2 (ed. K. Senneset), Balkema, Rotterdam, pp. 623–628.
- Biancotti, A. (1981) Geomorphologia dell'Alta Langa (Piemonte Meridionale). *Memori Societe Italiano Scienze Naturale, Milano*, **22** (3), 59–104.
- Boni, A. (1941) Distacco e scivolamento di masse a Cissone, frazione di Serravalle delle Langhe. *Geofisica Pura e Applicata*, **3** (3), 1–19.
- Brass, A., Wadge, G. and Reading, A.J. (1991) Designing a Geographical Information System for the prediction of landsliding potential in the West Indies, in *Neotectonics and Resources* (eds M. Jones and J. Cosgrove), Belhaven Press, London.
- Canuti, P., Casagli, N., Ermini, L. *et al.* (2004) Landslide activity as a geoindicator in Italy: significance and new perspectives from remote sensing. *Environmental Geology*, **45** (7), 907–919.
- Cortemiglia, G.C. and Terranova, G. (1969) La frana di Ciglié nelle Langhe. *Memorie Societa Geologica Italiana*, **8**, 145–153.
- Crist, E.P. and Ciccone, R.C. (1984a) A physically based transformation of Thematic Mapper data - the TM Tasseled Cap. *IEEE Transactions on Geoscience and Remote Sensing*, **GE-22**, 256–263.
- Crist, E.P. and Ciccone, R.C. (1984a) Application of the tasseled cap concept to simulated thematic mapper

- Murphy, W. and Vita-Finzi, C. (1991) Landslides and seismicity: An application of remote sensing. Proceedings of the Eighth Thematic Conference on Geologic Remote Sensing, Denver, Colorado, April 29–May 2, pp. 771–784.
- Polloni, G., Aleotti, P., Baldelli, P. *et al.* (1996) Heavy rain triggered landslides in the Alba area during November 1994 flooding event in the Piemonte Region (Italy), in *Landslides*, vol. 3 (ed. K. Senneset), Balkema, Rotterdam, pp. 1955–1960.
- Rengers, N., Soeters, R. and Van Weston, C. (1992) Remote sensing applied to mountain hazard mapping. *Episodes*, **15** (1), 36–45.
- Sacco, F. (1903) La frana di Sant'Antonio in territorio di Cherasco. *Annali – Reale Accademia di Agricoltura di Torino*, **46**, 3–8.
- Skempton, A.W. and DeLory, F.A. (1957) Stability analysis of natural slopes in London Clay. Proceedings of the 4th International Conference on Soil Mechanics and Foundation Engineering, 2, pp. 378–381.
- Taylor, D.W. (1948) *Fundamentals of Soil Mechanics*, John Wiley & Sons, Inc., New York.
- Tropeano, D. (1989) An historical analysis of flood and landslide events, as a tool for risk assessment in Bormida valley. Suolosottosuolo, Congresso Internazionale di Geoingegneria, Turin, Italy, 27–30 September, pp. 145–151.
- Varnes, D. (1984) Landslide hazard zonation: a review of principles and practice. Commission on Landslides of the International Association of Engineering Geology, United Nations Educational Social and Cultural Organisation, Natural Hazards, No. 3.
- DMAAC (1981) *Tactical Pilotage Chart of Algeria, Libya*, 1st edn, Series TPC, sheet H-3D, Scale 1:500,000, Defence Mapping Agency Aerospace Center, St Louis, MO.
- GPCC (1987, 1992–1993) Global Precipitation Climatology Centre homepage: <http://www.dwd.de/research/gpcc> (accessed 20 February 2009).
- Lee, H. and Liu, J.G. (1999) Spatial decorrelation due to the topography in the interferometric SAR coherence image. Proceedings of the International Geoscience and Remote Sensing Symposium (IGARSS'99), Hamburg, Germany, 26 June–2 July 1999, IEEE Press, Piscataway, NJ, vol. 1, pp. 485–487.
- Liu, J.G., Black, A., Lee, H. *et al.* (2001) Land surface change detection in a desert area in Algeria using multi-temporal ERS SAR coherence images. *International Journal of Remote Sensing*, **22** (13), 2463–2477.
- Liu, J.G., Capes, R., Haynes, M. and Moore, J.McM. (1997a) ERS SAR multi-temporal coherence image as a tool for sand desert study (dune movement, sand encroachment and erosion). Proceedings of the 12th International Conference and Workshop on Applied Geologic Remote Sensing, 17–19 November, Denver, Colorado, ERIM, Ann Arbor, MI, pp. I-478–I-485.
- Liu, J.G., Moore, J.McM. and Haigh, J.D. (1997b) Simulated reflectance technique for ATM image enhancement. *International Journal of Remote Sensing*, **18**, 243–255.
- Rudolf, B., Hauschild, H., Rueth, W. and Schneider, U. (1994) Terrestrial precipitation analysis: operational method and required density of point measurements, in *Global Precipitations and Climate Change*, NATO ASI Series I, vol. 26 (eds M. Desbois and F. Desalmond), Springer-Verlag, Berlin, pp. 173–186.
- Zebker, H.A. and Villasenor, J. (1992) Decorrelation in interferometric radar echoes. *IEEE Transactions on Geoscience and Remote Sensing*, **30** (5), 950–9.

## Section 21.4

Ahrens, C.D. (1994) *Meteorology Today: An Introduction to Weather, Climate, and the Environment*, West Publishing, St Paul, MN, pp. 514–521.



# 22

## Industrial Case Studies

This chapter describes two industrial case studies conducted by the authors in collaboration with other co-workers. Some of the issues surrounding these cases are highly confidential and so the material here has been confined to those aspects of the work which are directly related to remote sensing and GIS, and largely to the methodological aspects of the projects. In the first case (Section 22.1) the work has been carried out jointly with Mr Anders Lie of NunaMinerals A/S. The second case (Section 22.2) represents a small part of a wider project funded by UNICEF, and carried out for Gibb Africa Ltd, by Image Africa Ltd (UK) and Aquasearch Ltd (Kenya).

### 22.1 Multi-criteria assessment of mineral prospectivity, in SE Greenland

#### 22.1.1 Introduction and objectives

In this case study we describe the data and methodology used to enable a multi-disciplinary assessment of prospectivity for a number of economic commodities, namely nickel, copper and PGEs (Platinum Group Elements), in previously unexplored terrains of south-east Greenland. Since

this is a very large and ambitious project, we cannot do justice to its full complexity here and so this chapter contains a summary of the data preparation and methodological aspects, rather than the results which in detail are highly confidential. Some early, regional results, in the form of an example prediction map, are shown here since it is difficult to convey the concepts without visualizing the results to some extent.

This project is being conducted under a joint venture agreement between NunaMinerals A/S and Image Africa Ltd (UK). Its results are partly intended as a tool for attracting investors for the further development of any worthy areas identified during the project. The main reasons for conducting the project are explained in this chapter and lie in the fact that the south-east coast of Greenland comprises vast and unexplored Achaean and Proterozoic terrains which already have shown potential for hosting mineral discoveries. It is ideal for a GIS-based assessment using remotely sensed and other regional geoscientific data and so is a good case study example for this book.

The project involves three independent phases: firstly, the testing, compilation and assessment of rock and sediment sample material; secondly, the systematic, multi-parameter spatial analysis of remotely sensed and all available geoscientific data; and finally, the exploration and ground

validation of target areas pinpointed during the course of the project. The study area consists of two parts: the Reference (SW Greenland) and Survey (SE Greenland) areas, as illustrated in Figure 22.1. The general approach is to use the ‘fingerprints’ of known mineral occurrences in the Reference area to help predict new occurrences in the Survey area by identifying significant spatial patterns in the various datasets we have at our disposal. This approach is not new to exploration; the use of GIS to conduct this kind of regional spatial analysis is well documented elsewhere (Bonham-Carter, Agterberg and Wright, 1988; Knox-Robinson, 2000; Chung and Keating, 2002, to name but a few) but in Greenland this is a novel tactic.

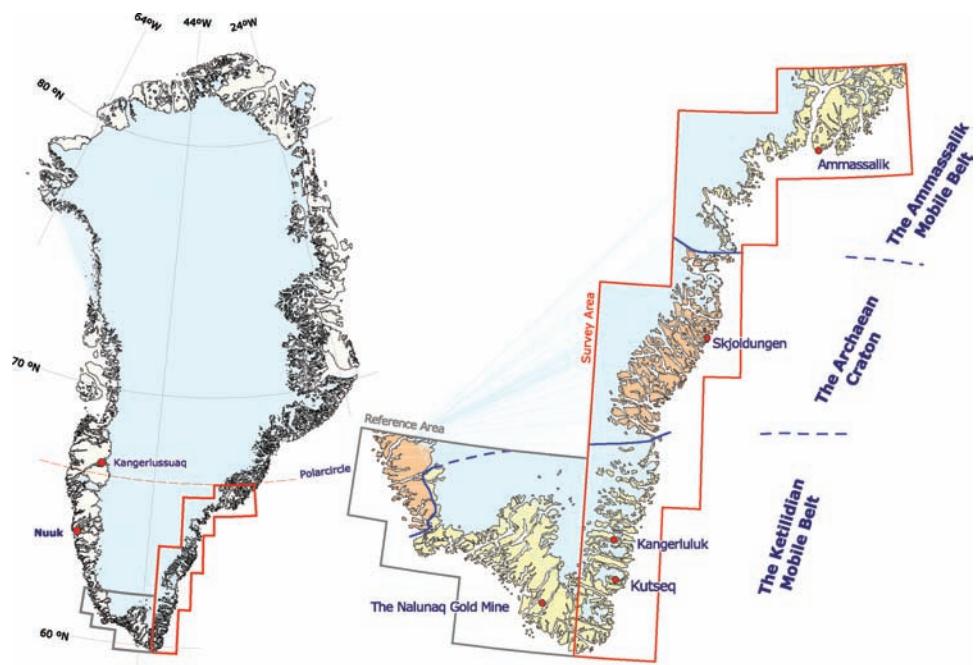
### 22.1.2 Area description

The project concerns 24 500 km<sup>2</sup> of highly exposed, poorly explored Archaean and early Proterozoic shield terrain which has significant inferred potential for hosting mineral deposits, and which stretches from Kap Farval in the south to the Ammassalik

Peninsula in the north. The terrain is extremely remote and only a handful of expeditions have been conducted to this part of Greenland (these are well documented). We have built on that knowledge during our work in the summers of 2006 and 2007, and visited some of the same locations, plus a great many more. Logistically, fieldwork is hampered by the persistent presence of icebergs along some sections of the coast, making passage by ship hazardous. Helicopter reconnaissance is hindered by the lack of any refuelling stations, meaning that fuel must be carried onboard ship. Fieldwork can be done by ship alone but it restricts accessibility to near-shore localities and means that far less distance can be covered. A mixture of the two is optimum.

Being in the ‘rain shadow’ of the Greenland ice sheet means that the east-coast terrain is generally dry and barren. Compared with the west coast, it is also steep, largely ice covered and almost devoid of vegetation and wildlife. The altitude and steepness further necessitate the use of a helicopter to conduct effective field reconnaissance and sampling work.

The Reference area is well studied and a great wealth of data and experience has been gleaned



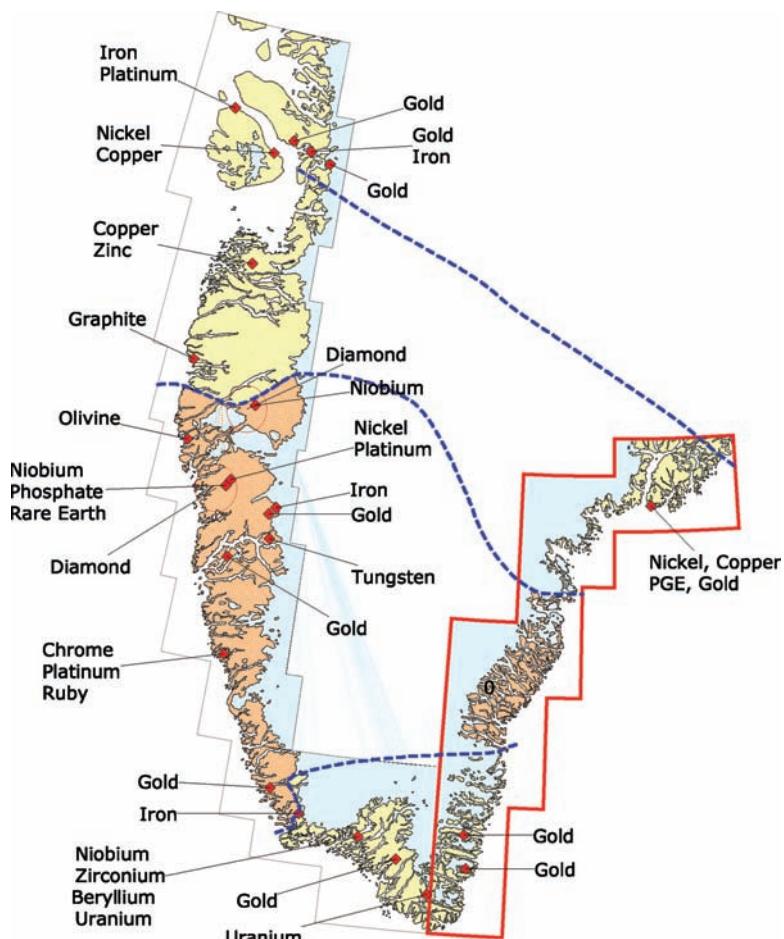
**Figure 22.1** (a) Map of Greenland showing the HMDP area; and (b) the project Reference and Survey areas

from it. The Survey area has, in contrast, largely been mapped only at regional scales. The geological understanding of the west coast exceeds that of the south-east coast and this may partly explain why very little commercial exploration has been conducted here. The map of known mineral occurrences in Greenland is testament to this, as it can be seen that there are far more on the west than on the east coast (Figure 22.2); it is well known that the potential of finding new occurrences is perceived to be greater where occurrences have already been found. The discrepancy in understanding, knowledge and mineral potential between west and east coast terrains is being addressed but to gain parity in such a large area, effectively and relatively quickly, a novel approach is necessitated.

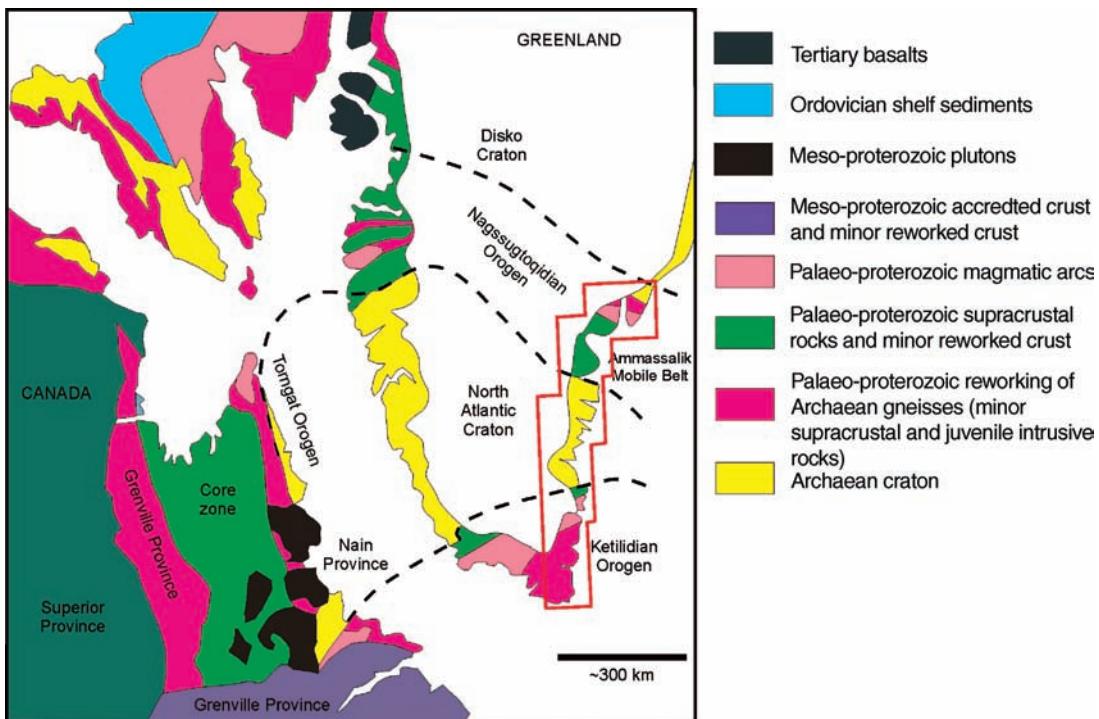
### 22.1.3 Litho-tectonic context – why the project's concept works

The Precambrian shield of south-east Greenland comprises three distinct basement provinces: Archaean terrain reworked during the early Proterozoic (Ammassalik Mobile Belt, of the Nagssugtoqidian Orogen), Archaean terrain almost unaffected by Proterozoic or later orogenic activity (the North Atlantic Archaean Craton), and juvenile early Proterozoic terrain (the Ketilidian Mobile Belt or Orogen). This project involves assessment of all three terrains.

It is thought that the Nagssugtoqidian Orogen extends beneath the Greenland ice cap to the east coast and that it can be closely correlated with the Torngat Orogen of north-eastern Canada (Van Gool



**Figure 22.2** Mineral occurrence map of Greenland with the project Survey area shown in red



**Figure 22.3** Proterozoic reconstruction of Greenland and Canada, showing the Atlantic–Arctic litho-tectonic trend. Modified after Van Gool *et al.* (2002). The approximate position of the Survey area is shown by the red polygon

*et al.*, 2002), see Figure 22.3. It is also thought that Greenland is closely related to similar, well-explored terrains in Finland and Russia. We conclude that the east coast has all the same litho-tectonic suites and litho-geochemical characteristics as the west coast and as provinces in Canada, and that it should therefore have similar potential to yield mineral deposits. This is our justification for exploring for a series of mineralization styles, which are already known to exist elsewhere.

used in the generation of prediction maps. These include komatiite-hosted nickel–copper–PGE, mafic–ultramafic intrusion-hosted deposits, lode gold and calc–alkali porphyry deposits. The deposit model focused on here is the komatiite-hosted nickel–copper–PGE type. To describe adequately and illustrate the methodology and results for all commodities and deposit models evaluated in this project would far exceed the scope of this chapter and, for the purposes of illustrating the methodology, is unnecessary.

#### 22.1.4 Mineral deposit types evaluated

This sizeable area also comprises many different litho-structural and geochemical settings; under this project, a number of well-known mineral deposit models are therefore being evaluated within these terrains. Each type has been characterized according to economic commodity, geological setting and pathfinder minerals (primary and secondary), to aid the selection of input layers

#### 22.1.5 Data preparation

##### 22.1.5.1 Published maps

Maps from the east coast are available at rather coarser scales than other parts of Greenland, with the exception of a few localities where detailed work has been undertaken by the Geological Survey of Denmark and Greenland (GEUS), and these have been made use of where possible. The Reference area

and the Lindenow Fjord area are covered by maps at a scale of 1 : 100 000 and these were also made use of. Maps at a scale of 1 : 500 000 formed the backbone of the extraction of geological background knowledge and were used to target the fieldwork during the summer 2007 campaign (two map sheets cover the Reference and Survey areas). These regional maps contain considerable internal geometric distortions, making it very difficult to georectify them accurately; this emphasizes the importance of remotely sensed imagery in providing an accurate base map for interpretation and data capture.

#### 22.1.5.2 *Lithology*

These published maps have been used to guide the interpretation and spectral discrimination, using ASTER imagery, of supra-crustal packages, mafic, ultramafic and alkali intrusions and generally any non-gneiss/granite outcrops. The absolute positions of outcrops captured from maps have then been corrected using the ASTER imagery.

For simplicity, four categories have been created and coded with integer values of 1 to 4. These are ultramafic rocks (4), basic meta-volcanic supra-crustal packages (3), gabbros and other basic igneous intrusions (2) and other alkali igneous intrusive bodies (1). Since the vast majority of exposed rocks in this part of Greenland are unminalized crystalline lithologies such as gneisses,

granite and granodiorite, we are really interested in identifying any other outcropping lithology. In the case of komatiite-hosted Cu–Ni–PGE deposits, the mineralizations occur in massive sulphide form and these are often small, dark and almost impossible to identify directly using remote sensing. We must therefore concentrate on identifying any potential hosts such as ultramafic intrusions and basic volcanogenic supra-crustal packages (which may also contain ultramafic rocks). Likewise, any mapped gabbroic intrusions should be included since they may be associated with ultramafic bodies which have not yet been mapped (given that mapping is not particularly detailed). We have not included the younger, ‘Gardar’ intrusive bodies since these are related to Atlantic opening events and are not significant for this deposit type. At the first stage of prediction, to avoid giving bias to one lithology type or another, we use a background value of 0 to represent gneiss/granite exposures and a value of 1 for all other lithologies (see Table 22.1). We also accept that the maps will not show all lithological outcrops of interest, and that some may have been incorrectly mapped, so we hope to detect others from the ASTER imagery.

#### 22.1.5.3 *Structure*

Structural features have also been extracted from the maps and since these are at a coarse scale, the

**Table 22.1** Numerical coding for thematic input layers to the spatial modelling of komatiite-hosted Ni–Cu–PGE prospectivity

Thematic layer	Class represented	Original values	Buffer distance (m)	Coded values
Lithology	Ultramafic bodies	4	—	1
	Basic intrusives (e.g. gabbros)	3	—	1
	Basic supra-crustal (meta-igneous)	2	—	1
	Alkali intrusives	1	—	1
	Crystalline basement (gneiss and granite)	0	—	0
Structure	Terrain boundary	3	10 000	1
	Major	2	500	1
	Minor	1	300	1
	‘Unfractured’ (massive) basement	0	0	0

features extracted tend to represent major faults and structural sutures; again, their positions are often inaccurate and require correcting or interpretation using the ASTER imagery. The vast majority of fractures and faults in the spatial modelling database have been interpreted from the ASTER imagery, working at a scale of about 1 : 50 000.

For the purposes of the spatial modelling exercise, these interpreted (and mapped) structures are taken to represent zones of fracturing (potentially including faulting, shearing, jointing and other planes or zones of weakness and discontinuity). These are considered important as conduits and potential destinations for mineralized fluids, which are otherwise too small to be directly detected in the remotely sensed imagery and may not be obvious during fieldwork. Since they are represented as linear features but are taken to represent wider zones in reality, buffers are calculated around the linear features at variable distances, dictated by the relative tectonic significance of the structure. For instance, major terrain boundaries are coded with the highest rating value and minor faults the lowest; these values are then used in producing the buffers with the largest buffer distance for the most significant structures and so on. The actual distances used were chosen on the basis of field experience and the published literature, and vary between 500 and 10 000 m. A fuzzy distance-rated buffer layer was then created in which buffer zones are coded from a value of 1, closest to the structure, decreasing gradually to 0 over the distance set by the structure's significance. The remaining, unfractured background is assigned a value of 0 (see Table 22.1).

#### 22.1.5.4 Mineral occurrences

Mineral occurrence data were derived from several Sources and their function is two-fold: to provide evidence data for the prospectivity predictions; and to provide data for the later cross-validation of those prediction maps (not shown in this case study since the work is incomplete and ongoing).

These data are simply point positions of known mineral occurrences but with no numerical value as such, except for nominal values to identify the commodity (or commodities) found at that location. The occurrences used here have been derived from the GEUS south-west Greenland database, from the Ujarassiorit public geochemical database (managed by the Bureau of Minerals and Petroleum, of Greenland) and from the laboratory results of our own fieldwork.

For the komatiite-hosted Ni–Cu–PGE deposit model a total of 78 known mineral occurrences (containing pathfinder elements for this deposit model) were used as evidence in the spatial modelling. The point localities were coded (as in Table 22.2), buffered at a distance of 300 m, to allow for inaccuracies in the positioning information, and then converted to a binary raster with values of 1 for occurrences and zero for the background.

#### 22.1.5.5 Remotely sensed image data

Some 60 ASTER level 1B scenes have been used and these formed the most reliable and accurate base-map framework for all other mapping and data capture; they were also used to generate a series of spectral indices.

**Table 22.2** Mineral occurrences used and their reclassified values

Mineral occurrence	Commodity represented	Original values	Coded values in binary raster
Ni, Cu	Dominantly nickel plus copper	1	1
Cu, Au	Dominantly copper plus gold	2	1
Au	Gold	3	1
Au, Cu	Dominantly gold plus copper	4	1
PGEs	Platinum Group Elements	5	1
Various	All other elements	6–99	0
—	Background	—	0

Ideal acquisition time is during the summer since the snow disappears in June and reappears in September, making July and August the ideal months; all the data used in this project were acquired in those months, between 2001 and 2006.

DEM<sup>s</sup> were first generated from each ASTER scene, using the onboard ephemeris data. Very limited ground control data exists (only at limited sites visited during scant field trips) and this is insufficient for any practical purposes here. The DEMs were then used to ortho-rectify each ASTER scene.

Prior to any image processing, the ASTER images were pre-processed to mask the very brightest targets, ice (and snow), and the darkest targets, water, from the data. Without doing this, the enhancement of geological targets becomes very difficult, if not impossible. Generally speaking, sea pixels are not difficult to remove when the water is clear, since its reflectance in the infrared is almost zero. The problem comes when water contains suspended rock flour, as is commonly the case in the upper reaches of fjords around Greenland. Reflectance from the suspended rock flour renders the water almost indistinguishable from that of land pixels. Defining a threshold at which to mask flour-loaded water becomes a very delicate operation, since some land pixels will be lost if great care is not taken. There will be cases where the removal of the last pixels, representing the most stubborn flour-laden waters, has to be done manually, using a vector mask (which is extremely tedious and to be avoided if at all possible). With ASTER data acquired on very different dates, the algorithm used to remove ice and sea required customizing for each set of illumination conditions. Fortunately, one long strip of data collected on a single day covered much of the central and southern part of the Survey area, making this job slightly less arduous. The logical expression constructed required the examination of the visible and thermal bands over land, sea and ice, to identify suitable thresholds; an example is shown in Figure 22.4.

**22.1.5.1 Image processing for general visualization** A standard false colour composite mosaic (of ASTER bands 321 RGB) of the entire area was constructed and used for both general visualization

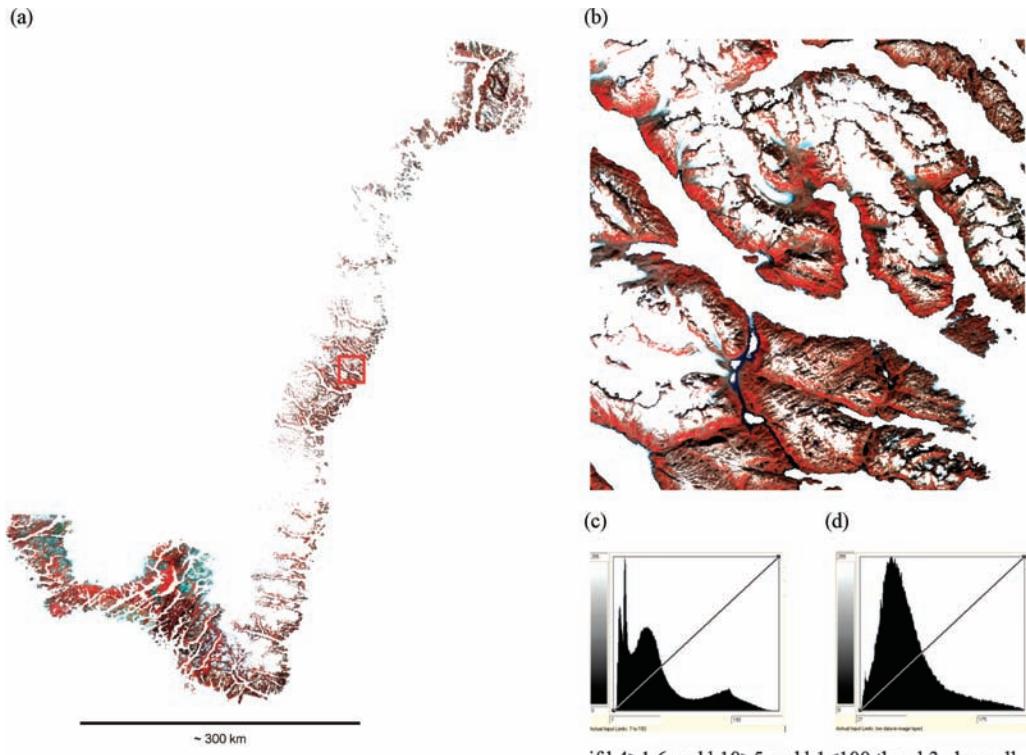
and interpretation, and also for field reconnaissance planning (shown in Figure 22.4).

For general visualization, this is best produced from data which still contain snow and the ice masses since it is easier to interpret this in the field. Individuals not so familiar in working with remotely sensed data often find it difficult to navigate using imagery when it is in its processed and rather abstract form. The best image for general visualization and navigation in the field should be as simple as possible so that image features can be readily correlated with real objects on the ground, (including glaciers and ice-capped mountains).

**22.1.5.2 Targeted spectral image processing** Ratio indices have been derived from ASTER VNIR, SWIR and TIR data to highlight various chemical characteristics of rocks and minerals, while also suppressing topographic shadowing. These indices, or relative absorption band depth images as they have been described (Crowley, Brickey and Rowan, 1989; Rowan, Mars and Simpson, 2005), use the DNs from the bands marking the shoulders of diagnostic absorption features; the principle is illustrated in Figure 22.5. Here they have been derived to highlight the spectral absorption features of iron oxides and hydroxides, in weathered gossaniferous zones ( $b2/b1$  and  $b4/b3$ ), MgOH or carbonate content ( $b6 + b9/(b7 \cdot b8)$ ) and silica paucity, to reveal mafic and ultramafic rocks ( $b12/b13$ ) and used to target lithologies of interest (where these are extensive enough to be detected).

Our primary mineralized targets are all associated with lithologies which are considerably poorer in silica content than the granites/gneisses that make up the vast majority of exposure. Many observed mineralizations have associated zones of hydrothermal alteration but in general these are of very limited extent and so undetectable by ASTER. So spectral indices aimed at highlighting hydrated minerals (as would be normal in detecting alteration and mineralization in other litho-tectonic settings) tend to reveal areas of weathered granite/gneiss rather than any alteration associated with mineralization.

Some of these are illustrated in Figure 22.6 for a small part of the Survey area. The indices suggest



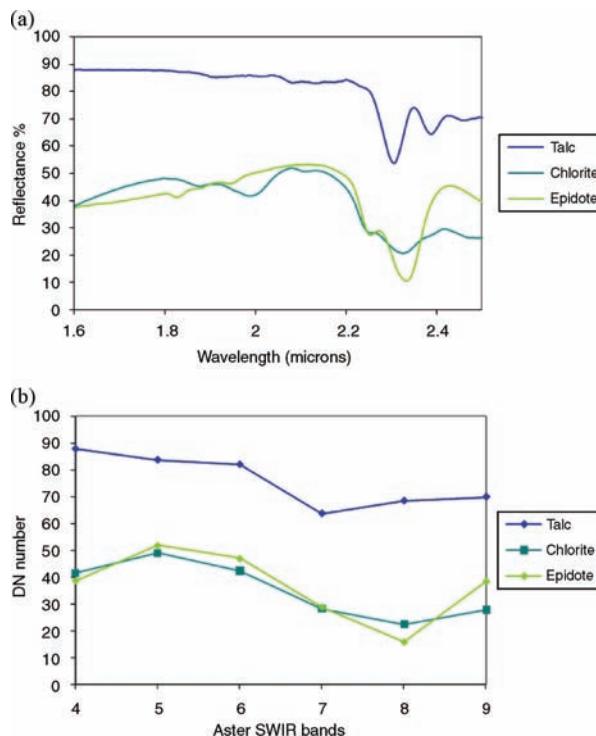
**Figure 22.4** Standard false colour composite mosaic of: (a) the entire area (Reference and Survey); (b) detailed area on the east coast (indicated by the red box in (a)), with ice and sea masked out; and (c) histograms for band 3 of this detailed area before and (d) after masking of ice and sea

that there are more outcrops of supra-crustal packages than the ones mapped; this is not surprising given the mapping scale. We also see that there is variation within the mapped packages: some zones are more mafic than others (Figure 22.6c) and some zones are altered (Figure 22.6d). The mafic silicates index in Figure 22.6c exploits the Restrahlen feature of reflectance spectroscopy at thermal infrared wavelengths and indicates the presence of silicate minerals which are characteristic of mafic igneous rocks rather than the acidic group; here red colours indicate the mafic igneous zones within the supra-crustal packages, against the acid crystalline background of gneisses (blue). The index shown in Figure 22.6c shows the intensity of absorption caused by MgOH bonds in hydrated (altered) basic igneous minerals, such as chlorite, epidote, talc, and may indicate where the supra-crustal packages

are intensely fractured, weathered or altered through fluid passage. Other useful indices include that highlighting oxidized iron at the surface, as an indication of gossan development, and that which highlights iron in silicate phase, as an indication of the presence of mafic igneous rocks (as shown in Figure 22.6d).

#### 22.1.5.6 Geochemical sample data

The geochemical data used here are a compilation of laboratory analysis results from three sources: rock samples collected in the 1960s, rock and sediment samples collected during two targeted field reconnaissance seasons in 2006 and 2007, and data from the Ujarassiruit database. Samples from the 1960s and those collected in 2006 were also analysed using a field spectrometer (PIMA and ASD, respectively), the results of which have been very useful in characterizing both the background



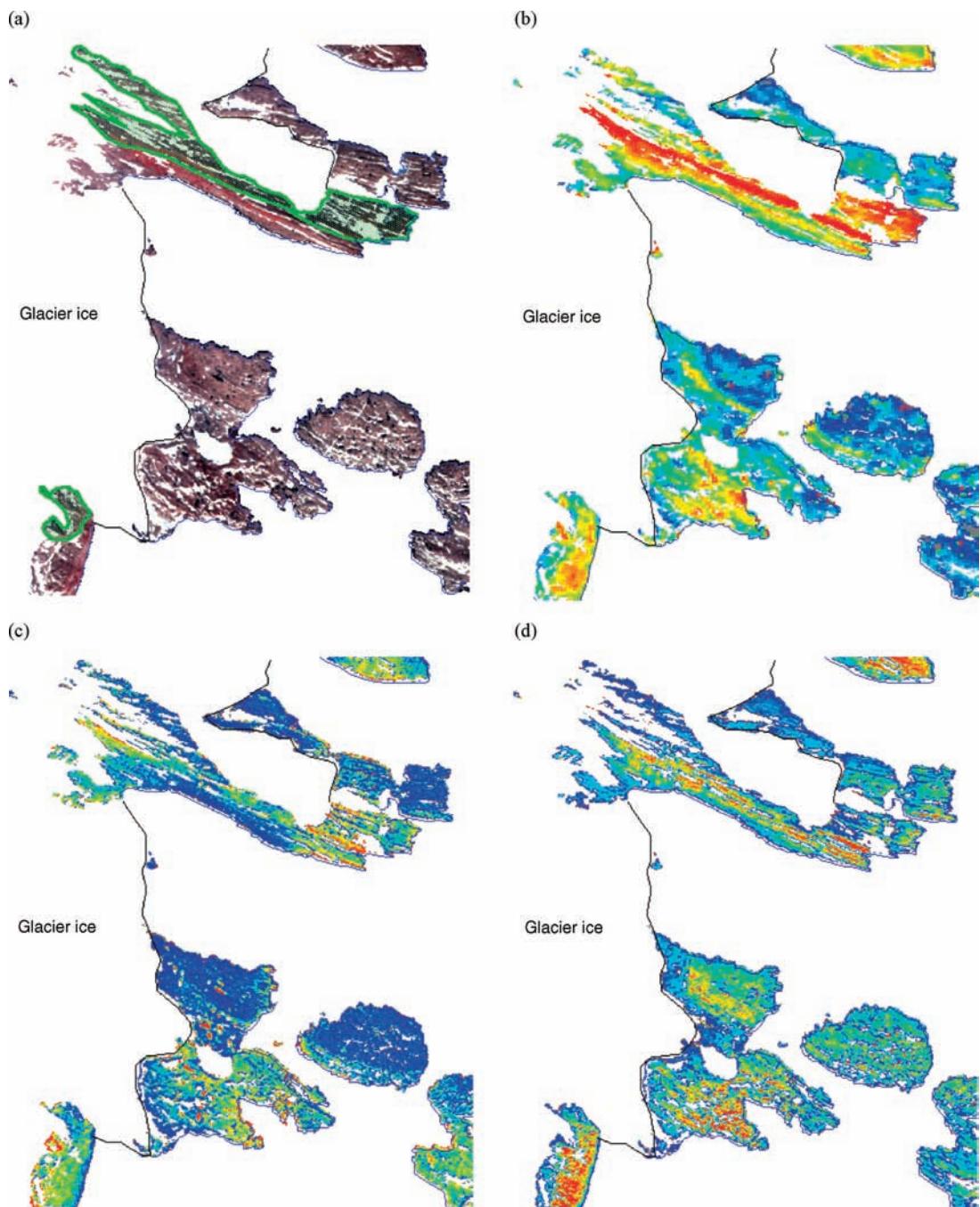
**Figure 22.5** The basis of spectral index (or band depth image) generation: (a) SWIR reflectance spectra of talc, chlorite and epidote; and (b) the same spectra convolved to ASTER SWIR bandwidths. The sum of the band DN at the maxima either side of the feature is divided by the sum of the band minima DN, in this case bands 6 and 9 over 7 and 8, will highlight the broad absorption feature between 2.2 and 2.4 microns which is characteristic of talc and especially chlorite and epidote

spectral signatures and those of altered samples from known mineralized localities.

The coordinate references of some of the data points are more reliable and accurate than others. This is particularly the case for the Ujarassiuut data where, in some cases, the positions were not obtained using a GPS instrument but were approximated on a map sometime after collection.

From these data, an ‘intelligent’ geochemical database has been developed, principally for the purposes of this project but also for best practice in managing company assets. The data have been categorized according to the method of analysis and of field collection, the project they are associated with, as well as when and by whom they were collected. In this way, each result can be viewed critically, and with understanding of the likely significance and reliability of its contribution to the modelling of prospectivity.

For each particular deposit model, certain pathfinder elements were selected: nickel, copper, cobalt, chromium, gold, platinum and palladium in this case. The sample point data for each pathfinder element were then interpolated, using a minimum curvature tension (spline) local polynomial function, to produce a continuously sampled representation of estimated element concentration over the entire area. For the elements used, the concentration values have been processed such that values below the detection limits (during laboratory preparation and analysis) were set to 0, and anomalously high values were clipped at an arbitrary value (of 1), and the logarithms of the remaining values were then used. This ensured that the interpolation was not biased either by extremely low values (which are of no interest) or by extremely high values (which may be produced by the nugget effect, and therefore are also of low interest). The use of log transformed



**Figure 22.6** Graah Islands area of south-eastern Greenland: (a) Standard false colour composite (ASTER 321 RGB) showing two supra-crustal packages as identified in published maps (green polygons); (b) mafic silicates spectral index ( $b_{12}/b_{13}$ ) displayed as a pseudo colour layer with increasing value from blue through red (0–255); (c) MgOH spectral index (e.g. epidote, chlorite, talc) indicating altered basic volcanic rocks, displayed with the same colour table as (b); and (d) ferric iron in silicates index (ferromagnesian minerals). The field of view is 14 km

values is standard practice in geochemical analysis and is appropriate here since it is the pattern (distribution) of values that is of interest rather than the absolute concentration values themselves.

The ‘gridded’ element data were then scaled to a 0–255 value range for input to the spatial modelling. Once again it is important that ice and sea are masked from the data before scaling to the 0–255 value range and not afterwards. If scaling is done afterwards, the value range of the significant pixels (i.e. land pixels) will be suppressed (see Figure 22.7). We are most interested in finding areas where the values are truly elevated with respect to the background values, rather than minor local perturbations in the background level. In this respect, it is also important to notice the shape and spread of the histogram. A histogram which shows a single, normally distributed population without skew (or with slightly positive skew) is highly likely to represent background levels of that element. What we would like to see is a negatively skewed histogram that has several peaks (i.e. representing multiple populations), as shown in Figure 22.7(d); see also Section 15.2.1.

A summary of the geochemical pathfinder layers used and input to the spatial prediction modelling, for this deposit model type, is shown in Table 22.3.

#### *22.1.5.7 Regional geophysical data*

Regional geophysical data have also been made available to the project, in the form of gravimetric point sample records which have been corrected for free-air and Bouguer anomalies. These data were then interpolated to produce a surface representing regional gravity for the project area. It was found that although these data could be used for relative interpretations of anomalies, they could not be used within the spatial modelling without introducing a considerable regional bias, since they had not been corrected for the presence of the Greenland ice sheet which exerts a very broad effect on the data.

Since we have insufficient information to make a rigorous correction of this effect, we perform a ‘makeshift’ correction by subtracting a regional average, as calculated using a  $31 \times 31$  average filter, from the individual values, thus removing the low-frequency variation (caused by the ice) and leaving the high-frequency information representing anom-

alies of geological origin (see Figure 22.8). The result can then be used within the spatial modelling since the data have effectively been normalized: high values in one area have similar absolute value range, and similar meaning, to those in another location, a vital consideration with any data used in this kind of work. Once again, the interpolated, ice-corrected gravity data must be masked for ice and sea before being scaled to an 8 bit (0–255) value range, for the same reasons outlined previously.

#### **22.1.6 Multi-criteria spatial modelling**

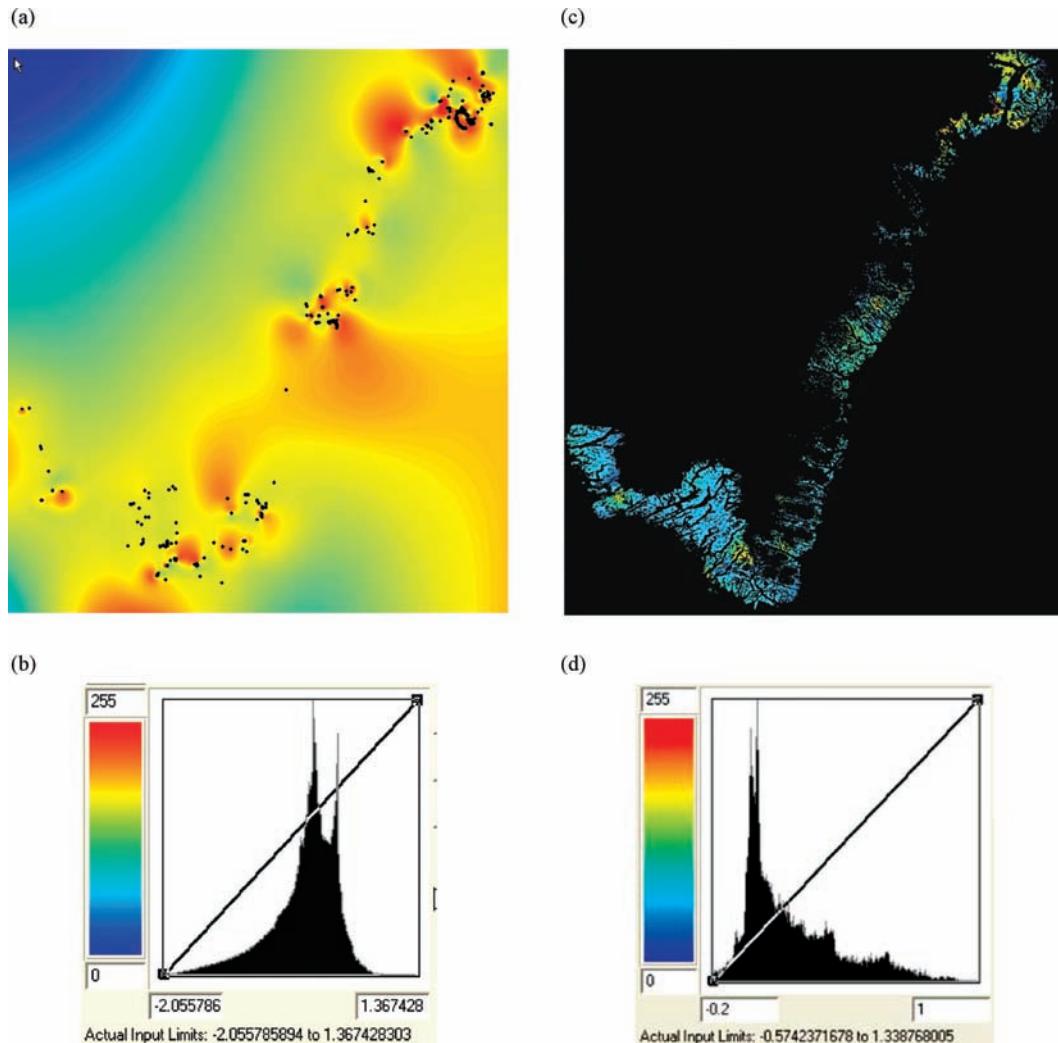
Broadly, the strategy is to conduct a semi-automated suitability mapping of the Survey area. The spectral, structural, geochemical and stratigraphic characteristics of proven mineral occurrences in the Reference area of south-western Greenland have been used as evidence from which to estimate the probability of discovering comparable mineral occurrences in the Survey area of south-eastern Greenland.

The Spatial Modelling and Prediction (SPM) software package (first developed by C.F. Chung) is being used here for the evaluation of our criteria, and to generate probability estimates (or prediction maps). Once the semi-automated prediction maps representing the potential for identifying new targets have been produced satisfactorily, they will be weighted by market-driven factors, such as commodity price and accessibility, to identify economically favourable areas for further detailed ground validation.

##### *22.1.6.1 Method of layer combination*

All data must have been normalized to a common scale for use in the multi-criteria analysis, and in this case a byte scale (0–255) has been used. They must also be coded so that all are positively correlated with high suitability. No input evidence must be contradictory or the results will be meaningless.

Bayesian probability discriminant analysis was chosen as the most appropriate method for combining the input layers and estimating prospectivity (as described in Section 18.5.3). The main reason for doing so is that this method involves the least subjective decision making during criteria combination. Some subjectivity cannot be



**Figure 22.7** (a) The interpolated Ni tenor grid of entire area, with sample points (black dots), and its histogram (b); and (b) the interpolated grid of the land area and its histogram (d), after masking sea and ice. The histogram in (d) is skewed towards the low end of the value range and clearly shows several significant populations, indicated by the smaller peaks on the right. This represents the effective removal of background information and bias caused by the estimated low and high values in the ice and sea areas

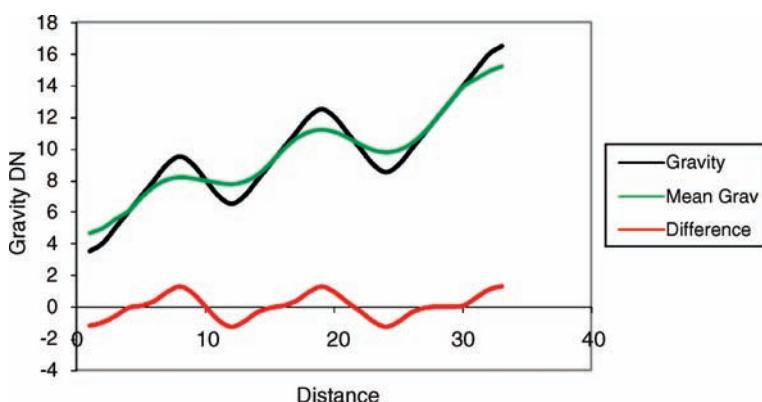
avoided along the way, in the capture and preparation of the data, but at the point where the input layers are combined we have to acknowledge that we do not have sufficient knowledge to apply any further rules or weighting of the input layers. In effect, we want to let the data speak for themselves and tell us what actual patterns exist. This necessity is a natural reason for choosing Bayesian probability as the method. We use known occur-

rences as our evidence, to gain the ‘signature’ (distributions) in the input layers, and use these to make predictions of the likelihood of finding new occurrences.

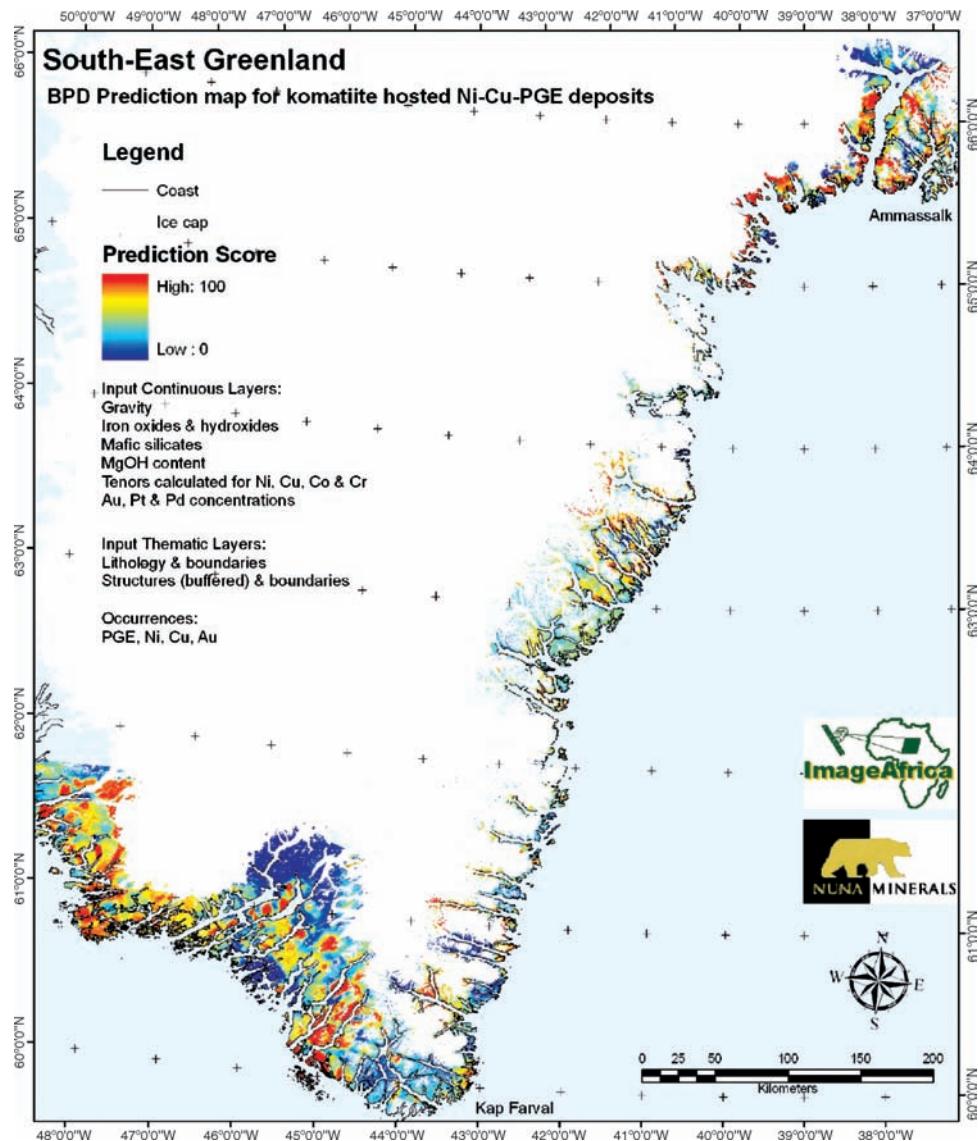
If known mineral occurrences are in fact associated with high values in the input layers, then the output prediction map should predict them. If so, then we should have more confidence in the map’s prediction of other areas as showing potential.

**Table 22.3** Numerical coding for continuously sampled input layers to the spatial modelling of komatiite-hosted Ni–Cu–PGE prospectivity

'Continuous' layer	Class represented	Method of processing/ preparation	Coded value range
<i>ASTER:</i>			
Mafic content	Absence of free silica (quartz)	TIR (b13/b10)	Increasing values, on an 8 bit value range (0–255)
Iron oxides	Oxidized iron at the surface ('rusty zones')	VNIR (b2/b1)	
MgOH	Presence of MgOH in mafic minerals	SWIR (b6 + b9)/(b7 + b8)	
<i>Geochemical:</i>			
Tenors	Interpolated from nickel, copper, cobalt and chromium tenors, i.e. metals in sulphide phases	Rock samples only	
Au, Pt and Pd	Interpolated gold, platinum and palladium content	Rock and sediment samples	
<i>Geophysics:</i>			
Gravity	Corrected for regional effect of Greenland ice sheet	Interpolated from regional Bouguer and free-air corrected ground sample point data	



**Figure 22.8** Schematic representation of the ice correction applied to the gridded Bouguer-corrected gravity surface values. The black line illustrates both the local variations and the regional trend imposed by the ice cap (here sloping upwards left to right), the green line represents the mean (low-frequency aspect) of this surface (as calculated using a spatial averaging filter), and the red line represents the differential surface (high-frequency aspect) which reveals the significant fluctuations in gravimetric potential without the regional bias imposed by the large mass of the polar ice cap



**Figure 22.9** Prediction map representing prospectivity for komatiite-hosted Ni-Cu-PGE deposits in the Reference and Survey areas

One result for the deposit model described here is shown in Figure 22.9. It reveals a series of high-value areas, some more extensive and more intense than others, at various locations throughout the Survey area. Some of these we have confidence in, others we suspect to be exaggerated by the overwhelming presence of a few high gold values when this is used as a pathfinder element; in addition, there are some areas that we think should appear more significant than they do here. Clearly further

modifications are required. The path to achieve a satisfactory result is not a linear or easy one, but one where criticism is necessary, at every stage.

#### 22.1.7 Summary

Data quality is one of the most significant issues here and probably in all cases of this type. A vital lesson is that criticism of and understanding of the

data, rather than blind acceptance, is absolutely crucial. Time must be spent during the data preparation stages, and long before allowing the data to be input to the multi-criteria spatial analysis; the old adage of ‘rubbish in, rubbish out’ is very true here.

There are several, very tangible reasons for conducting this project in south-east Greenland, and for conducting it in the most correct way possible.

In comparison with the rest of Greenland, the depth of knowledge in the Survey area is far lower than elsewhere. The published maps are at a far coarser scale here than anywhere else. It is the only part of Greenland’s coast which has never been colonized and where the logistics of travel and work have prevented all but a very few, highly focused expeditions. As a consequence, this project now represents a significant contribution to the collective knowledge base for the south-east coast of Greenland.

We chose the Bayesian probability method of multi-criteria evaluation since it is one of the most objective. Given the vast complexity of the area and the volume of data involved, we wanted to let the data speak for themselves, and prevent bias, as much as possible, at least in the way they are combined. The Bayesian predictive discriminant analysis method used here identifies the signature of known occurrences in each of the input data layers and then assesses the probability of finding occurrences at other locations on the basis of the combinations of the input evidence layers (as described in Chapter 18). Some bias is unavoidable since we do not have unlimited data and the size of the area means we must spatially sub-sample the data to make the processing manageable.

The purpose of conducting this at a very regional scale of observation and prediction is to demonstrate potential in a few target areas. These few key areas will then become the subject of further, more focused predictions. This is the logical way to proceed and the way that all studies based on remote sensing generally proceed. It would be pointless to tackle the entire area at the highest conceivable level of detail from the start. The cost of this would be astronomical and would be extremely inefficient since we actually expect that the vast majority of the project area will be shown to be non-prospective – just one of the drivers for this work is to prevent

ourselves and others from wasting valuable time and funds in those areas.

Further work will include repeating this method for other mineral deposit types. These will necessitate the use of different spectral indices, path-finder geochemical elements, lithological units and mineral occurrence layers, as appropriate to each model type. In each case, we have a certain expectation of where the most predictive locations are likely to be, on the basis of our own field experience. In other words, the ground ‘truthing’ is partly provided by our own observations, even though we have only visited a tiny fraction of the area. If our result succeeds in predicting the locations that we know to be prospective, our understanding is reinforced and we can have more confidence in the map’s ability to predict previously unknown locations as being prospective. We will later use cross-validation curves to back up our own intuitive reaction in a more quantitative way. These will give an indication of the effectiveness of our prediction methodology, and which of our input layers and occurrences are most significant; they may even reveal unexpected anomalies in the data, occurrences or results which are significant in some previously unknown way.

We conclude that:

- The size of the project area requires a novel and effective methodology to be successful.
- The paucity of data compared with other areas of Greenland means that the level of understanding here needs to be raised rapidly, to catch up.
- This part of the Greenland coast represents a geologically, tectonically, prospectively and topographically very complex area – one in which poor data quantity and quality, and little collective ground experience, mean that the potential for error in this type of work is enormous. As rigorous an approach as possible is therefore necessary.
- This effort represents both the most likely way to achieve some success and a significant investment for future exploration in this part of Greenland. It is, in short, the right thing to do in this context.

Despite our best efforts, there are some false positives in the results, and because of data paucity, there will almost certainly be some localities

that are poorly predicted. Here our own experience and knowledge will be required to rule out the former, and a little imagination will be needed to pick out the latter. This stage of work is only the beginning. In the next phase of the work, it may become necessary or desirable to use different methods of analysis to derive prediction images. Perhaps at such a time, we may have clearer ideas of the specific geological controls involved and can therefore employ more rigid decision rules or have more selective control over uncertainties. We hope that several of the areas predicted in this work will form the focus for the next stage of exploration, data collection and more detailed prediction.

## Acknowledgements

This work has been funded by NunaMinerals A/S and by Image Africa Ltd, as a joint venture project. We are grateful to the GEUS for the provision of the regional geophysical data, and to Dr Bo Muller Stensgard for his valuable technical advice on the use of the spatial prediction modelling software in this context.

## 22.2 Water resource exploration in Somalia

### 22.2.1 Introduction

This case study is based on work conducted as a collaborative effort between Image Africa Ltd and Aquasearch Ltd, for Gibb Africa Ltd, and was funded by UNICEF. The work involved the evaluation, through remote sensing, of the region around the city of Hargeisa, in Somalia (formerly British Somaliland), within a broader project whose remit was to identify potential bulk water resources for this city. Various criteria for such potential resources had to be satisfied within this remit. For instance, any suitable bulk resource should be no more than 100 km from the city, and it should ultimately be capable of supplying a demand of 40 000 m<sup>3</sup>/d.

The general approach was to conduct a relatively detailed desk study, in two phases, followed by some

targeted fieldwork. The first phase consisted of regional-scale geological interpretation and identification of target areas which warranted more detailed investigation. The second phase comprised similar but more focused interpretation of those target areas, using two- and three-dimensional visualization, using DEMs generated during the project, as well as spectral enhancement via image processing. The results of phases 1 and 2 image interpretation and digital mapping were carried out in parallel with, and then combined with, a detailed literature search and evaluation of borehole and other available hydrological data. The detailed nature of the two desk study phases was especially significant given the hostilities which were at the time preventing completion of the work through field investigations, a situation which persists to this day. The work presented in this case study represents a summary of the first and second phases, as an illustration of the approach taken for a case of this type.

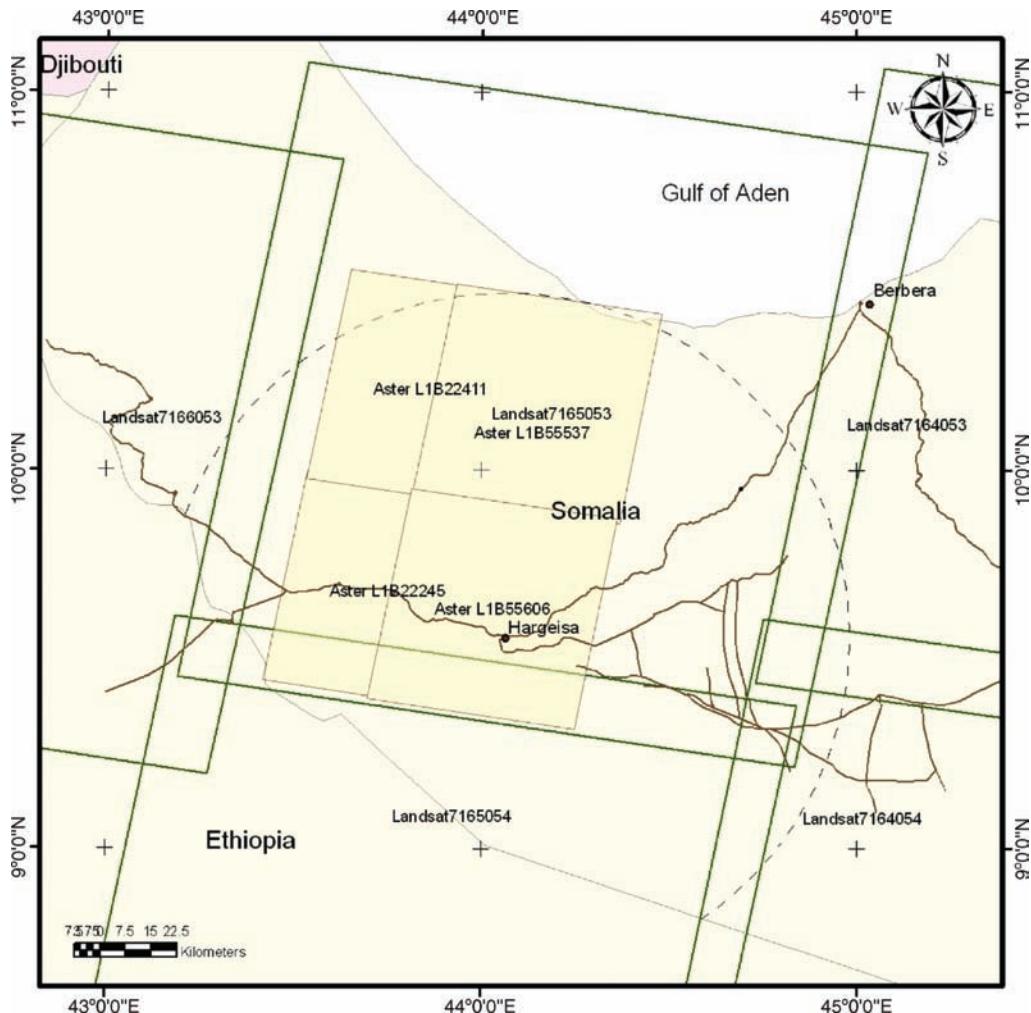
Specifically, the objectives of this part of the work were:

- To provide regional geological interpretation and mapping.
- To provide evidence of the location and extent of potential aquifers (as described by hydrogeologist M. Lane of Aquasearch Ltd).
- To make recommendations on areas for further investigation.

The first objective would be satisfied largely using medium-resolution data, in the form of Landsat-7 ETM+ imagery, an SRTM DEM, published geological maps and mean annual rainfall (MAR) data. The second and third objectives would involve the use of ASTER imagery and DEMs in addition to geological maps.

The study area occupies some 300 km<sup>2</sup> along the northern coast of Somalia (formerly British Somaliland), in north-east Africa, and is illustrated in Figure 22.10. It includes coastal alluvial plains in the north, volcanic escarpments, mountainous massifs, a limestone plateau and extensive desert sands to the south, and encompasses varied climatic conditions. The average rainfall here is 380 mm per year and this falls largely on the high ground.

There are a number of potential aquifers in this region. The most significant ‘extensive aquifers’



**Figure 22.10** Map of the study area in Somalia, NE Africa, showing the main cities and roads, the 100 km radius and the image scene coverage used in this work (Landsat-7: five clear, labelled polygons; ASTER: four yellow, labelled polygons). The 100 km radius marker and country border are used for scale in all regional images shown here

with the potential to yield bulk water resources could be found in any or all of the solid geological units. These units are the Precambrian fractured crystalline metamorphic rocks, Jurassic limestones and sandstones (Adigrat Formation), the Nubian Cretaceous sandstone and the Eocene Auradu limestone. The second most significant comprise local structural basins which are dominantly filled with Jurassic limestones and younger sediments. Then there are minor sedimentary aquifers, several of which were identified in previous studies in the western, central and eastern coastal plains respec-

tively. The alluvial and colluvial materials of the coastal plains were also considered.

### 22.2.2 Data preparation

The datasets used in this study are listed in Table 22.4. The Landsat ETM+ images used for this work cover an area of about  $300 \times 200 \text{ km}^2$ , span more than one climatic zone and were acquired during April, May and September of three different years. In such circumstances, the production of a

**Table 22.4** Summary of digital data used in this study

Data	Path/row	Acquisition dates
Landsat-7 ETM+	164/053–054, 165/053–054, 166/053	1999, 2000, 2001
ASTER	Four scenes	2002
ASTER DEMs (15 m)	Four scenes	2002
Maps	British geological 1 : 50 000 map sheets	Sheets 20–23, 23–34
Rainfall data	Mean annual rainfall	

seamless image mosaic is almost impossible. Images involving combinations of bands 1, 2 or 3 also suffer from the effects of atmospheric haze, and this is most noticeable in the areas closest to the Sahara. Atmospheric quality over the mountainous and coastal areas, by comparison, is very good. The haze effects are most noticeable therefore in the southern and south-eastern Landsat scenes, and are responsible for the poor contrast matching (and rather obvious scene boundaries) between these two and the northern and western scenes. Images involving bands 4, 5 and 7, in contrast, are affected less by the effects of haze, so that the scene boundaries are less noticeable in the image shown in Figure 22.16. In each of the image maps presented, the data have been contrast enhanced to match as closely as possible.

The MAR data were supplied in point form by Gibb Africa Ltd, and were gridded and contoured for correlation with other data. The SRTM DEM data were downloaded, as compressed 1° tiles (from the NASA public data distribution centre, see Appendix B.3) decompressed, georeferenced and ‘mosaiced’ into a single DEM of the entire study area, with a pixel size representing 90 m on the ground. DEMs (gridded at 15 m) were also generated from each ASTER scene, using between 6 and 20 control points identified in each scene. The easting and northing coordinate values were collected for these points, from the ortho-rectified Landsat ETM+ imagery, and the corresponding height values for these points were collected from the SRTM DEM. Each of the four ASTER DEMs was generated separately and then mosaiced. The discrepancy in elevation between any of the four scene DEMs was less than 5 m, providing an almost

seamless DEM mosaic of the target areas which is of sufficient accuracy to support mapping at 1 : 50 000 scale. Image data coverage is illustrated in Figure 22.10.

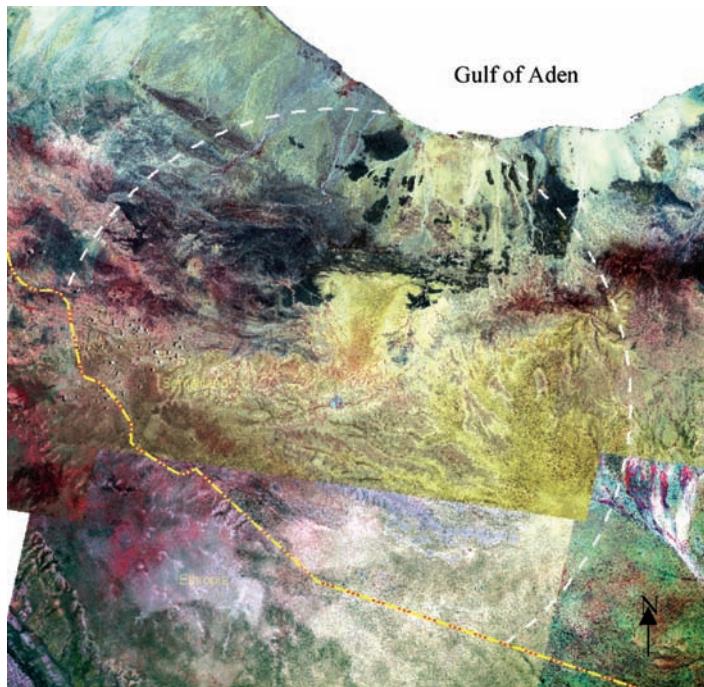
### 22.2.3 Preliminary geological enhancements and target area identification

The first phase comprised a regional-scale image interpretation and desk study, to identify major areas of interest and to extract as much information on regional geology as possible, given that map information was limited. This involved the use of broadband, medium-resolution, wide-swath imagery, Landsat-7 ETM+ in this case, to produce a regional geological interpretation, from which potential target areas were identified for further investigation. In addition, the SRTM DEM was used to extract regional structural information and to examine areas of aquifer recharge with respect to topography.

#### 22.2.3.1 General visualization and identification of natural vegetation and recharge areas

Once again we begin with a general visualization of the area, using simple colour composite images (true colour and standard false colour) and a shaded-relief image generated from the regional SRTM DEM.

The Landsat 432 colour composite reveals that the distribution of vegetation is not regular or even but is extremely sparse except in the western and central-eastern mountains (Figure 22.11). Calculation of an NDVI from the Landsat data and then



**Figure 22.11** Landsat 432 standard false colour composite image

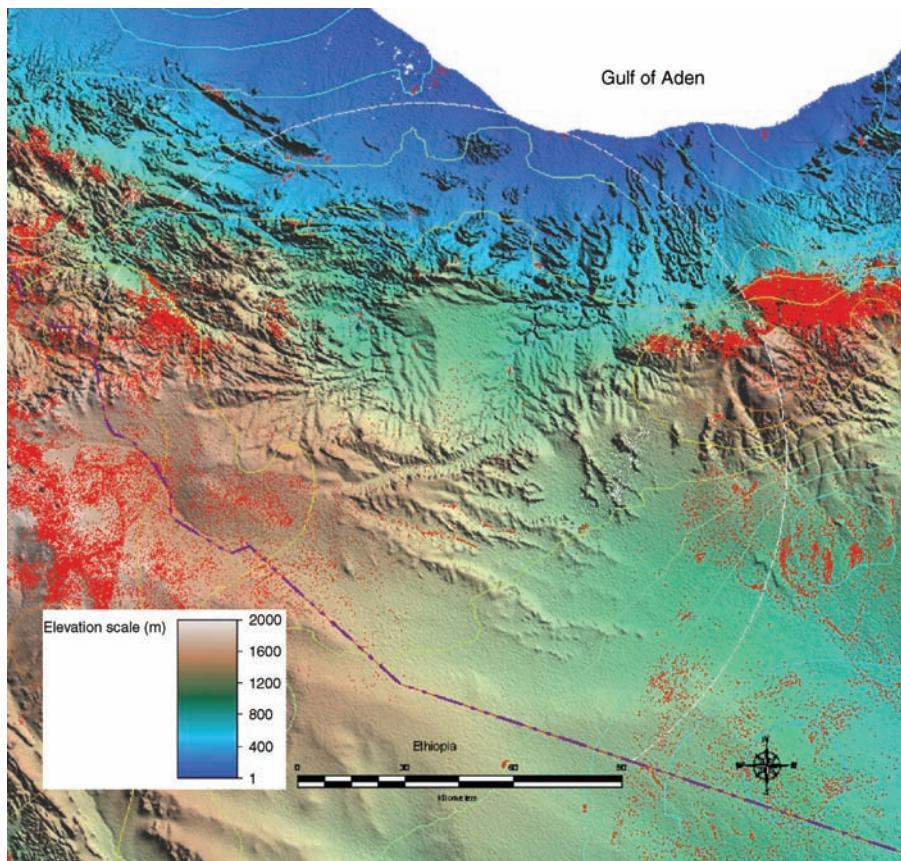
comparison of this with the MAR data and the DEM shaded-relief image (Figure 22.12) reveals that there is close correlation between the main centres of incident rainfall, persistent natural vegetation and high elevations. These areas comprise the principal areas of groundwater recharge for this region; they lie around the outer part of the 100 km zone around Hargeisa. Some small patches of vegetation can be seen in the valleys which cut the Haud Plateau, and at the base of the scarp slope of this plateau, to the south of Hargeisa itself.

#### 22.2.3.2 Regional geological interpretation

Simple colour composites, band ratios and the SRTM DEM were integrated and used to produce a regional geological interpretation, which was suitable for use at about 1 : 50 000 scale. Composites of bands 531 and 457 (Figure 22.13a and b) formed the backbone of this work as they provided very good discrimination of metamorphic basement lithologies, carbonates, volcanics and superficial sedimentary cover. The DEM and images together

allowed a very detailed structural picture to be produced (Figure 22.13c), especially in the area to the south of Hargeisa, where the Nubian and Auradu sediments are faulted and dip away to the south beneath the desert sands. Structure in the metamorphic basement massifs is too complex for interpretation at this scale of detail, so only the major terrain units were identified. A number of major sedimentary basins were identified within these massifs in the western part of the area, and these were predominantly filled with carbonate sequences.

The broad bandwidths of Landsat SWIR bands mean that carbonates and hydrated minerals (and vegetation where present) can be highlighted but not distinguished from one another. If we use a ratio of bands 5 over 7, we will highlight hydrated minerals (clays) and carbonates. By looking at the spectral profiles of the major units of interest, we find that the carbonates have, in general, higher DN values, in the daytime thermal infrared of Landsat band 6. If we modulate the  $b5/b7$  ratio with the TIR, we produce an index image which highlights all the major carbonate bodies in the area, in addition



**Figure 22.12** SRTM DEM shaded-relief map with thresholded NDVI (red) and mean annual rainfall contours overlain, to indicate the close correlation between rainfall recharge areas and natural vegetation distribution

to some gypsiferous deposits which are interbedded with them. The resulting image is shown in Figure 22.14; here the main targets are all highlighted and circled to identify them.

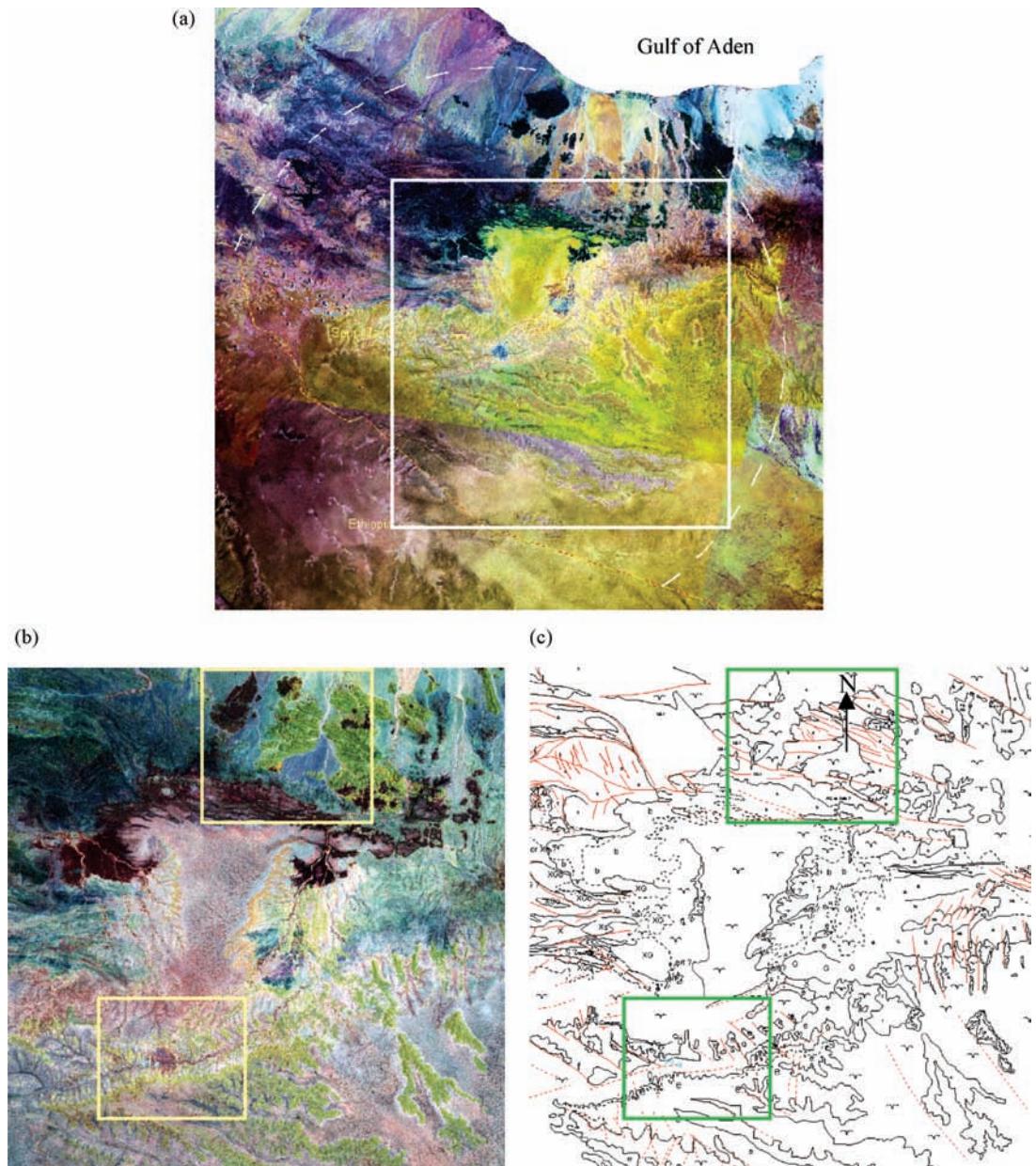
#### 22.2.4 Discrimination potential aquifer lithologies using ASTER spectral indices

The second phase involved sub-regional-scale interpretation of ASTER data, to evaluate the target areas identified in phase 1 and to assess whether these warranted further, detailed field-based investigations.

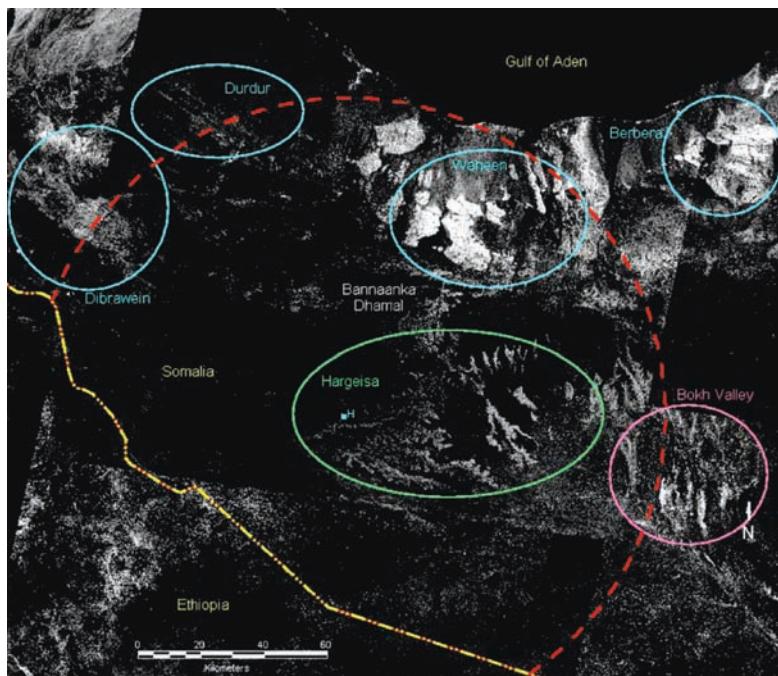
Given the lithological nature of the targets identified in phase 1, mainly carbonates and closely associated sandstones, ASTER is particularly suited

to the task of discriminating them. Using ASTER data, wavelength-specific variations enable the identification of carbonates (limestones, marble, calcite and dolomite), silicates (quartz, quartzites and different families of igneous rocks) and MgOH-bearing minerals and rocks (weathered, metamorphosed and altered basic igneous rocks), in addition to iron–aluminium silicates and iron oxides/hydroxides (both as alteration and weathering products), different families of hydrated minerals (clay alteration products, weathering breakdown products and evaporates such as gypsum and anhydrite) and of course vegetation.

The main target areas for investigation in phase 2 included the Bannaanka Dhamal, Agabar Basin and Nubian/Auradu contacts south of Hargeisa, the Waheen and the Dibrawein Valley. We cannot describe all the identified target areas within this



**Figure 22.13** (a) False colour composite mosaic of Landsat bands 531 DDS of the entire study area for general interpretation of geological features: sands from the Sahara appear in yellow and brown tones, with crystalline basement lithologies in dark blue and purple tones, and coastal alluvial sediments in the north in a variety of similar colours according to their provenance. The white box indicates the extent of the images shown in (b) and (c). (b) False colour composite of bands 457 DDS showing carbonate rocks in greenish tones, basaltic volcanics in dark reds, desert sands in pinkish tones and crystalline basement lithologies and coastal alluvial materials in shades of blue. (c) Detailed regional geological interpretation made from the Landsat imagery (lithological boundaries in black, structures in red). The two boxes in (b) and (c) indicate the locations of the detailed areas shown in Figures 22.16–22.18. Field of view is 340 km. The location of Hargeisa is indicated as H (cyan)



**Figure 22.14** Ratio of Landsat bands ( $(b5/b7)b6$ ), revealing carbonates and possibly gypsiferous rocks which form the target areas for detailed field-based investigation in phase 2: Jurassic limestones (cyan); Eocene and Cretaceous limestones (green); transported gypsiferous debris in the Bokh Valley (pink). The 100 km circle and political borders (red and yellow dashed lines) are also shown for reference

chapter so we focus on two of them: (a) the Hargeisa and (b) Waheen areas.

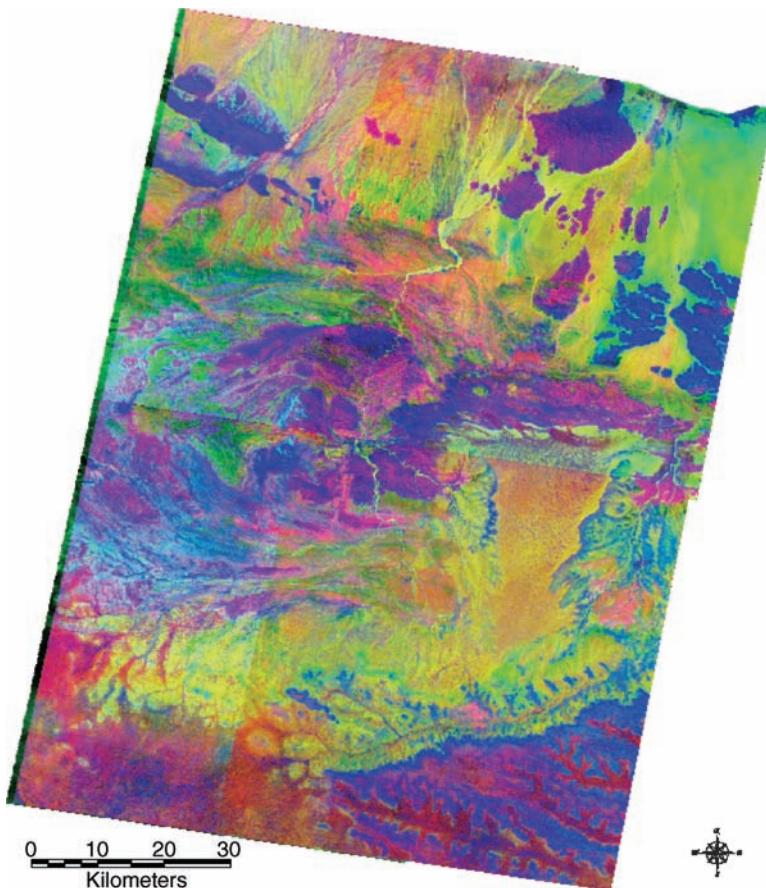
#### 22.2.4.1 Hargeisa area: Auradu limestone and Nubian sandstone

The objectives here are to enhance and delineate the outcrop of the main potential aquifers: the carbonates and sandstones; specifically the Eocene Auradu limestone and the Cretaceous Nubian sandstone lying stratigraphically below it. The Nubian sandstone is the principal aquifer for the whole of eastern Africa, its outcrops covering several million square kilometres – it represents the largest fossil aquifer in the world. It has been described as a hard ferruginous sandstone with several clay-rich intercalations.

Given the two distinct targets of carbonates and sandstones, we use ASTER VNIR, SWIR and TIR indices (relative absorption band depth images) to identify iron oxides, carbonates (limestones and marbles) and silica-rich materials (sandstones, granites, gneisses). Producing a composite result, bringing all these three indices together, in which

iron oxides/hydroxides (Fe–O and Fe–OH absorption) are enhanced and displayed in the red silicate-rich materials (S–O absorption, by quartz sands, gneisses, etc.) in green, and carbonates (C–O absorption) or weathered and altered basic volcanics (MgOH absorption) in blue. The result is a very brightly coloured image, shown in Figures 22.15 (the entire area), 22.16d (Hargeisa area) and in 22.18d (Waheen area).

In Figure 22.15, the desert sands of the Sahara in the south appear very noticeably in red tones. In the central area and to the north-west of Hargeisa, the yellow colours indicate high silica content and relatively high iron oxides; here the basement is composed of crystalline gneisses and granites and the surface may have a covering of desert sand or some desert varnish. The bright greens in the northern part of the area represent almost pure quartz sands of the Waheen and coastal alluvial systems. The basic Aden volcanics appear in magenta and bluish purple, indicating MgOH absorption of weathered (or altered) basalt and iron oxide/hydroxide absorption.

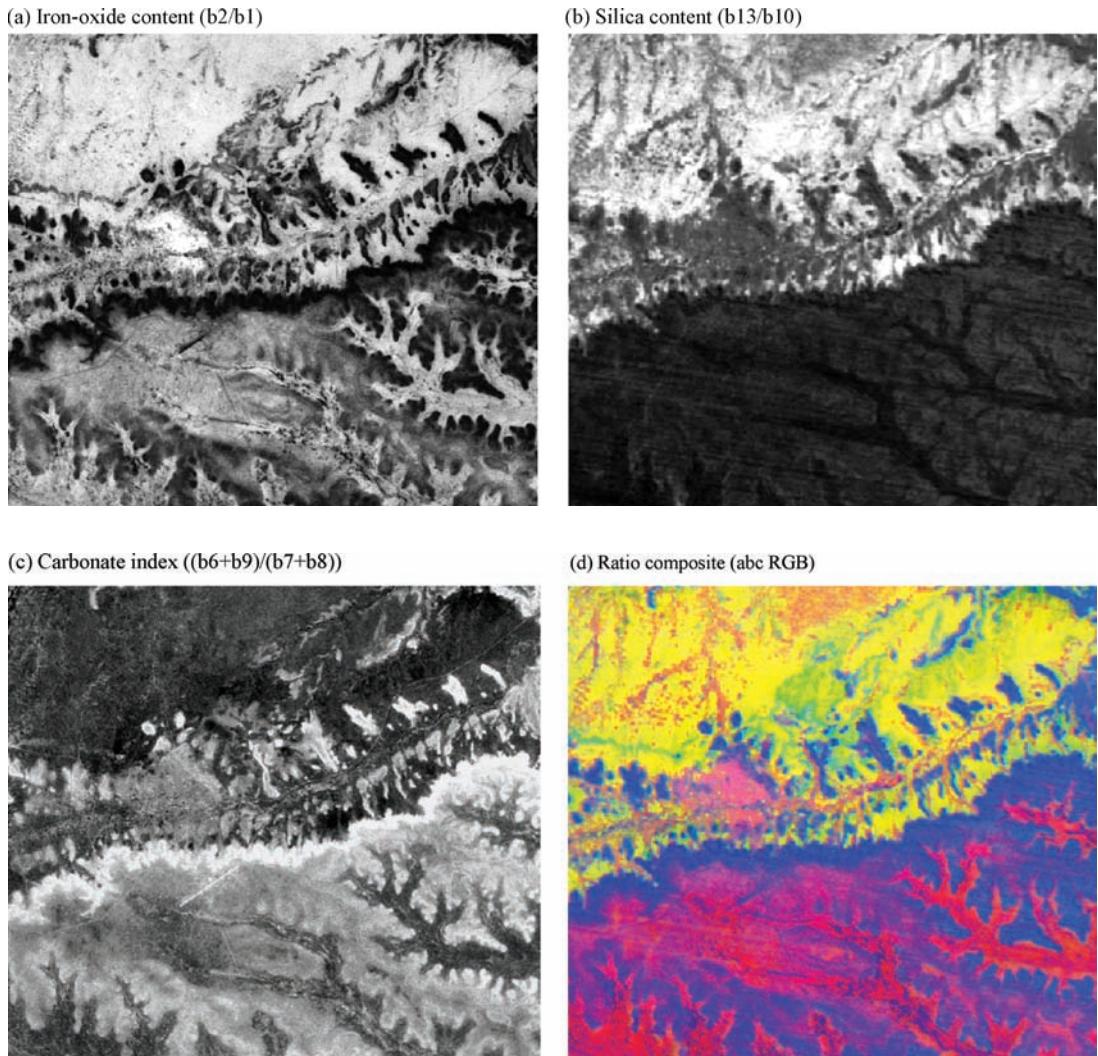


**Figure 22.15** Spectral index colour composite: iron oxides (red) highlighting red desert sands and ferromagnesian-rich lithologies, silica abundance (green) highlighting quartz sands and siliceous crystalline basement lithologies, and carbonate and MgOH abundance (blue) highlighting limestones and basic volcanics so that the basic volcanics appear in magenta whereas the limestones (and marbles) appear as pure blues, the coastal alluvial sediments appear in greens, oranges and reds whereas the limestones covered with sands appear cyan green and oxidized iron in sands appear orange

The remaining bright royal-blue tones represent the carbonates of the Waheen (north), the Dibrawein Valley (west) and Auradu limestones (south).

The second objective here in this area is to identify the Nubian sandstone specifically, if possible. The images in Figure 22.16 suggest that silica-rich targets cover the entire northern part of this area, so identification may not be straightforward. The 468 false colour composite in Figure 22.17a reveals the Auradu limestone in yellow tones, with the crystalline basement of the north and other areas around Hargeisa appearing in pale blue tones. There is also a narrow pink unit striking west to north-east, which in this band

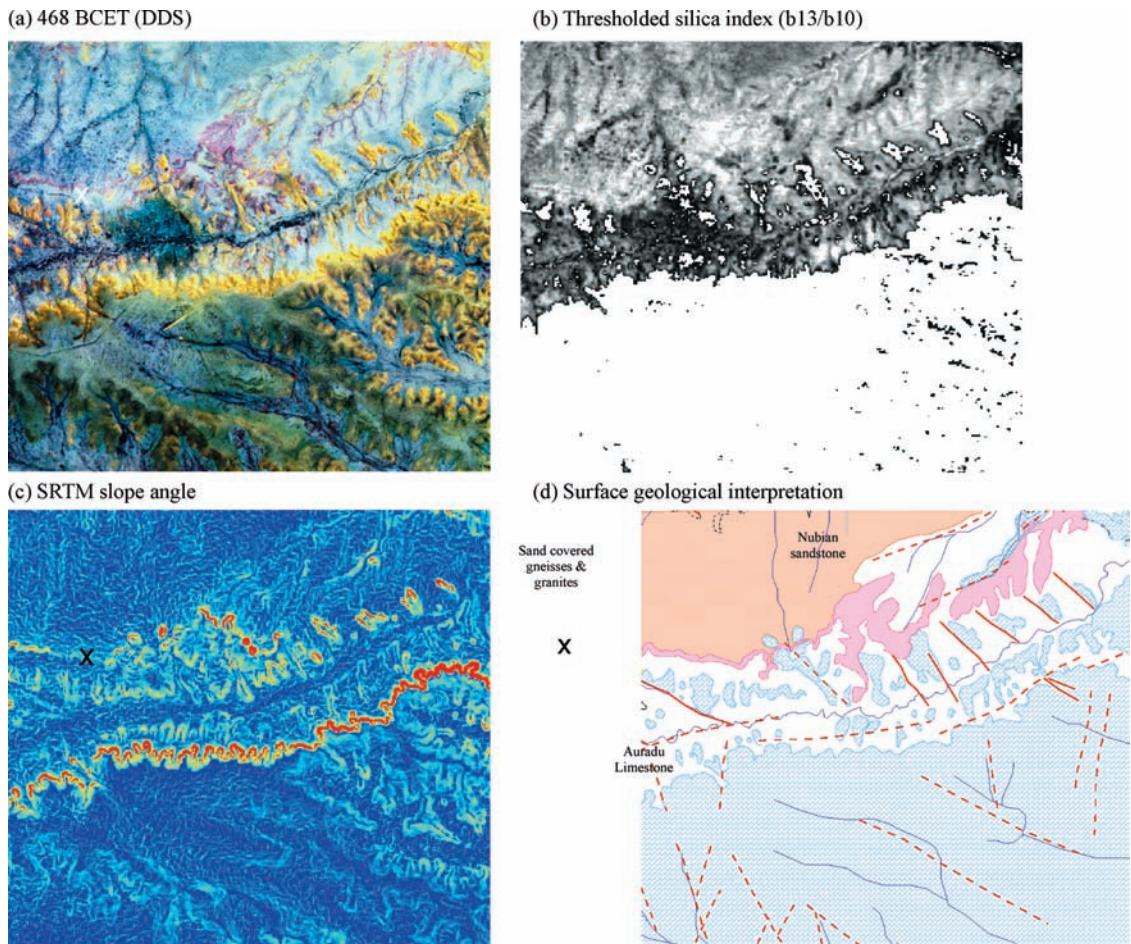
combination indicates the presence of hydrated minerals. Looking at the silica index shown in Figure 22.16b, we see that it has two distinct populations, a silica-rich and a silica-poor one. If we threshold this image to remove silica-poor areas, and stretch the remaining silica-rich population, the image in Figure 22.17b is produced. Here we can see that the areas of very pale blue in Figure 22.17a show up as having the highest silica content. The city of Hargeisa appears as an area of low silica values. We know that the Nubian sandstone outcrops in this area but cannot easily distinguish it from other silica-rich materials on the basis of spectral information alone.



**Figure 22.16** The Hargeisa area: (a) iron oxide content ( $b2/b1$ ); (b) silica content ( $b13/b10$ ); (c) carbonate index ( $((b6+b9)/(b7+b8))$ ); and (d) ratio composite image produced by combining (a), (b) and (c) as RGB: desert quartz sands appearing yellow and orange, with a contribution from high silica content (green) and iron oxides (red); crystalline gneissic basements (beneath the sands) appear in yellow (high silica); the Auradu carbonates appear in bright blue tones, some patches of green appear in valleys where carbonate and silica signatures are mixed. The upper surface of the Auradu limestone plateau, and the valleys cut into it, appear red suggesting it has a strong iron oxide content, perhaps produced by a surface coating or ferricrete. Field of view is 33 km in width

If we turn to the ASTER DEM of the area we see a number of breaks in slope running across the image, to the south and the north of Hargeisa. If we calculate the slope angle from the DEM (Figure 22.17c), we see these breaks in slope very clearly and notice that they coincide with the scarp slopes of the dipping sedimentary units here. A subtle scarp slope

can be seen in the west, which coincides with the narrow pink unit visible in Figure 22.17a; this can be followed eastwards but its outcrop pattern becomes more complex and it becomes more difficult to discern. The Nubian sandstone formation is reported to contain a number of clay-rich (perhaps mudstone) intercalations (Sultan



**Figure 22.17** The Hargeisa area: (a) 468 BCET (DDS) reveals the Auradu limestone in yellow tones; the silica-rich basement and sands to the north appear in pale bluish tones; a narrow pink unit ( $\times$ ) is visible running west to north-east and this colour indicates a lithology with a high clay content (the city is visible as a dark blue patch in the left centre). (b) The silica index ( $b13/b10$ ) thresholded and then stretched to reveal the silica-rich population representing the Nubian sandstone formation, crystalline basement lithologies and desert sands in the north of this image. A slope angle image (c) shows breaks in slope. The only clue to the Nubian sandstone lies in the topography. Between the pink layer and the yellow limestone sits the Nubian sandstone (white area) but its spectral properties are so similar to other silica-rich lithologies that it is indistinguishable from them. (d) Simple geological interpretation. The field of view is as in Figure 22.16

*et al.*, 2004); we suggest that the narrow pink band described previously represents the base of the Nubian sandstone formation.

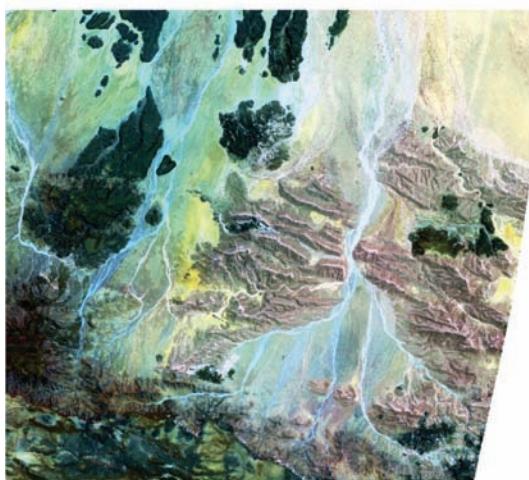
This example reveals the need for complementary datasets and field verification in conducting this type of work; the spectral information gives an honest indication of what is at the surface but the actual explanation for it may be elusive until other data can be introduced and/or a visit made to the site.

#### 22.2.4.2 Waheen: Alluvial gravels

The objective here is to establish any indication of water stored in the localized alluvial basins along the coast. The images indicated that the deposition sediment (transport) direction is broadly to the north and the Gulf of Aden.

The apparent absence of vegetation is revealed in Figure 22.18a. The materials here are granular and composed of silica-rich material (as shown in

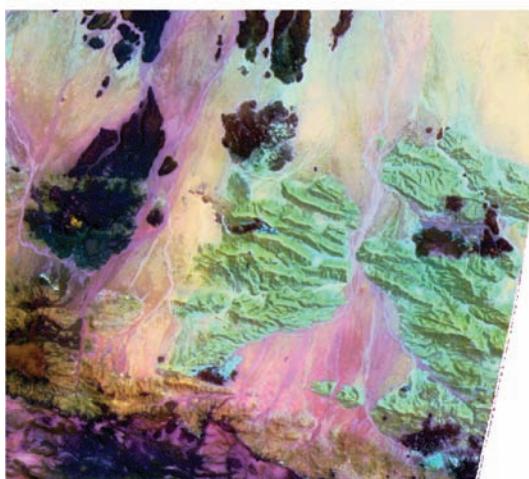
(a) 321



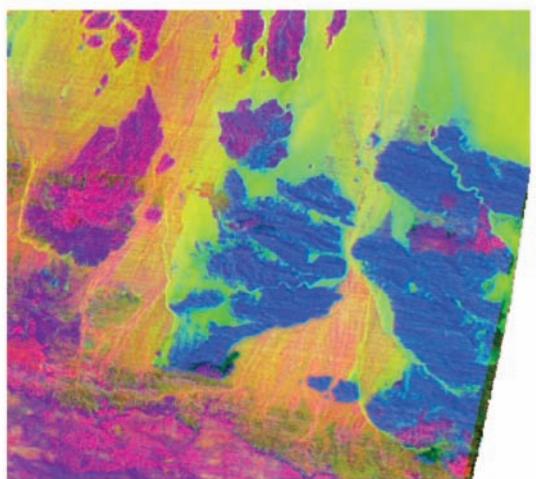
(b) NDVI



(c) 942



(d)



**Figure 22.18** Waheen region: (a) 321 BCET DDS colour composite revealing the apparent absence of verdant vegetation; (b) NDVI indicating that some vegetation exists on the central outcrops and on the basement massif in the south-west (NDVI values range between 0 and 0.1); (c) 942 BCET DDS false colour composite revealing central carbonates in green tones, volcanics in very dark blues and reds, metasediments in brown tones and alluvial gravels in pale pinks and yellows; and (d) ratio colour composite (R, iron oxides; G, silica content; B, carbonate/MgOH)

Figure 22.18d). These two indicators suggest either that there is little or no water stored in these alluvial gravels, or that any water here is at a depth too great to support any plant growth. There are other similar pockets of alluvial material along the coast. Some show distinct signs of vegetative growth and persistence but these lie at too great a distance from Hargeisa. The standard false colour composite (Figure 22.18a) suggests that no verdant vegetation exists in this area but a faint reddish colour in the

south-western corner of the area suggests some upland plant growth. The NDVI shown in Figure 22.18b suggests that some vegetation exists on the central outcrops and on the basement massif in the south-west but is absent from the alluvial plains around these outcrops.

The wadis that meander across the alluvial plains appear completely barren, supporting no vegetation, which suggests that any groundwater is too far below surface level to support any plant growth. In

either case, exploration for water in this region would appear to be futile. The only vegetative growth in this area is restricted to the south-west, on an escarpment formed by the Aden volcanics (magenta tones in Figure 22.18d).

The carbonates here are spectrally slightly different from those of the Auradu (shown in Figures 22.16 and 22.17), which are of Eocene age, and more closely resemble the Jurassic limestones which outcrop further to the north-east. They contain gypsum-rich units (the latter appear in cyan colours in the 942 composite image in Figure 22.18c); this and their spectral character suggest that they are of Jurassic age.

### 22.2.5 Summary

The phase 1 work allowed several areas to be identified as having potential for more detailed work. These were Bannaanka Dhamal, the Bokh Valley and the Dibrawein Valley. Several smaller, structurally controlled basins containing recent and Jurassic sediments and alluvium were also identified but many proved too small and at too great a distance from Hargeisa to represent any practical potential.

The difficulty in distinguishing the Nubian sandstone formation from other quartz-rich materials illustrates the need to look at other information. Image processing can always provide at least a partial answer, but something else will be necessary to complete it and this may, again, involve fieldwork.

The phase 2 work allowed the following to be achieved:

- Confident discrimination of the Auradu limestone outcrops in the Hargeisa region and to the east and west of Bannaanka Dhamal, which had not previously been possible.
- Enhancement of quartz-rich solid and drift deposits in the Hargeisa and Bannaanka Dhamal regions.
- Enhancement of vegetation and soils at 1 : 50 000 scale.
- Confirmation of the nature and influence of faulting on the outcrop geometry around the Hargeisa region, which may affect local aquifer existence.

Sadly, phase 3 of this work was never completed because of the political unrest in this country. More unfortunately, this project also revealed that the main remit of finding bulk water resources to supply Hargeisa was highly likely to fail. This work suggested that although potential aquifers exist, they are either too small or too far from the city to provide a viable economic supply. This project does, however, serve to illustrate a sensible strategy for other similar work.

## Questions

### Section 22.1

- 22.1 What are the principal sources and types of uncertainty and bias in any result?
- 22.2 What is the next logical step in the refinement of this project?
- 22.3 At the more detailed stage of observation, which other criteria combination method(s) would you choose? Critically evaluate your choice(s).

### Section 22.2

- 22.4 What recommendations would you make for the development of this work, given no budgetary or political restrictions?
- 22.5 Is it physically possible to distinguish between the different types of quartz (silica)-rich materials encountered here, spectrally or texturally? If so, how?
- 22.6 Having discriminated carbonates and basic volcanics (i.e. materials which exhibit CO or MgOH absorption in the SWIR), what other potential criteria could be used to differentiate between them?

## References

### Section 22.1

- Bonham-Carter, G.F., Agterberg, F.P. and Wright, D.F. (1988) Integration of geological datasets for gold exploration in Nova Scotia. *Photogrammetric Engineering and Remote Sensing*, **54** (77), 1585–1592.

- Chung, C.F. and Keating, P.B. (2002) Mineral potential evaluation based on airborne geophysical data. *Exploration Geophysics*, **33**, 28–34.
- Crowley, J.K., Brickey, D.B. and Rowan, L.C. (1989) Airborne imaging spectrometer data of the Ruby Mountains, Montana: mineral discrimination using relative, absorption band-depth images. *Remote Sensing of Environment*, **29**, 121–134.
- Knox-Robinson, C.M. (2000) Vectorial fuzzy logic: a novel technique for enhanced mineral prospectivity mapping, with reference to the orogenic gold mineralisation potential of the Kalgoorlie Terrane, Western Australia. *Australian Journal of Earth Sciences*, **47**, 929–941.
- Rowan, L.C., Mars, J.C. and Simpson, C.L. (2005) Lithologic mapping of the Mordor, NT, Australia ultramafic complex by using the Advanced, Spaceborne Thermal Emission and Reflection Radiometer (ASTER). *Remote Sensing of Environment*, **99**, 105–126.
- Van Gool, J.A.M., Connelly, J.N., Marker, M. and Mengel, F. (2002) The Nagssugtoqidian Orogen of west Greenland: tectonic evolution and regional correlations from a west Greenland perspective. *Canadian Journal of Earth Sciences*, **39**, 665–686.

## Section 22.2

- Crowley, J.K., Brikey, D.W. and Rowan, L.C. (1989) Airborne imaging spectrometer data of the Ruby Mountains, Montana: mineral discrimination using relative absorption band-depth images. *Remote Sensing of Environment*, **29**, 121–134.
- Rowan, L.C., Mars, J.C. and Simpson, C.J. (2005) Lithologic mapping of the Mordor, NT, Australia ultramafic complex by using the Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER). *Remote Sensing of Environment*, **99**, 105–126.
- Sultan, M., Manocha, N., Becker, R. and Sturchio, N. (2004) Analysis and modeling of paleodrainage networks recharging the Nubian aquifer Dakhla and Kufra sub-basins revealed from SIR-C and SRTM data. *Eos, Transactions, American Geophysical Union, Joint Assembly Supplement*, **85** (17).

# Part Four

## Summary

Rather than simply repeat the key points from the preceding chapters, our intention here is to present more of our personal views on essential image processing and GIS techniques for remote sensing applications. We may sometimes reinforce key points that have already been made but, in doing so, we hope to convey our thoughts in a wider sense, beyond the strictest technical aspects of the book.



# 23

## Concluding Remarks

### 23.1 Image processing

1. Despite the fact that our presentation of essential image processing and GIS techniques is technique driven, the use of these techniques in their working context (within a remote sensing project) should always be application driven. In a real application project, effectiveness and cost, rather than technical complexity, will dictate your data processing and we find that the simplest method is usually the best.
2. Image processing can never increase the information provided in the original image data but the use of appropriate image processing can improve visualization, comprehension and analysis of the image information for any particular application. There is no magic trick in image processing which will create something that does not already exist in the image, but it can be magical when image processing makes subtle things seem obvious and distinctive.
3. An image is for viewing! Always view the image before and after image processing: there are many fancy techniques to provide a so-called ‘quantitative’ assessment of image quality but your eyes and brain are usually far more accurate and more reliable.
4. Never blindly trust ‘black box’ functions in any image processing system. These functions use default settings and, for general purposes, are convenient and usually produce quite good results, but they are unlikely to produce the best results in more specific applications and are subject to unknown information loss. It is far more important to understand the principle of an image processing function than to master the operation of the image processing software package. In fact, only from a sound understanding of the principles can you fully explore and test the potential of an image processing system to its limits.
5. Colours and grey tones are used as tools for image information visualization. Digital images can be visualized in grey tones, true colour, false colour and pseudo colour displays. When using multi-spectral image data, beginning with simple colour composites is always recommended.
6. Our eyes are capable of recognizing many more colours than grey levels and it is therefore sensible to display a black and white image as a pseudo colour image to get the most from it. Strictly speaking, though, a pseudo colour image is no longer a digital image but an image of symbols: once the incremental grey levels are assigned to different colours, the sequential, numerical relationship between these grey levels is lost.
7. No matter which techniques are applied, the final step of displaying the resultant processed image will always be contrast enhancement, to stretch or compress the image to the dynamic

- range of the display device for optimal visualization. The histogram provides the best guidance for all image contrast enhancement.
8. Digital images are often provided in unsigned 8 bit integer type but, after image algebraic operations, the data type may change to real numbers. Always check the actual value range after such processing for appropriate display and correct understanding of the physical meaning of the derived images.
  9. Image algebraic operations are very versatile. Four basic arithmetic operations ( $+$ ,  $-$ ,  $\times$ ,  $/$ ) enable quite a lot of image processing; more complicated algebraic operations can, of course, offer more but it is important not get lost in an endless ‘number-crunching game’. In using multi-spectral image data, spectral analysis of the intended targets is essential in guiding your image processing towards effectiveness and efficiency.
  10. We usually achieve enhancement of target features by highlighting them but enhancement does not necessarily mean highlighting; it can also be achieved by suppression. For instance, a black spot on a white background is just as distinctive as a white spot on a black background. On the other hand, in a complicated and cluttered background, bright features are indeed more eye catching.
  11. Filtering is introduced here as a branch of image processing but it is, in fact, embedded in the very beginning of image acquisition. Any imaging system is a filtering system: for instance, a camera is a typical optical Fourier transform system. A scene of the real world is an assembly of spatial variation in frequencies from zero to infinity. Infinitely high frequencies can only be captured by lenses of infinite aperture. An optical imaging system of limited physical size can only offer a particular spatial resolution by filtering out the detail corresponding to higher frequencies than its aperture allows.
  12. In being a neighbourhood process, filtering enhances the relationships between the neighbouring pixels but the relationship between a filtered image and its original image is not always apparent. An understanding of the principle and functionality of a filter is necessary for mastery of its use and for the proper interpretation of its results. For instance, both gradient and Laplacian are generally regarded as high-pass filters but the functionality of these two types is quite different.
  13. The convolution theorem links the Fourier transform, in the frequency domain, to convolution, in the spatial domain, and thus establishes direct image filtering via a PSF (Point Spread Function) without the computer-intensive FT and IFT. In a wider sense, convolution is just one form of local window-based neighbourhood processing tools. Many ‘box’ kernel filters are based on this loose concept and flexible structure but do not necessarily conform to the strict definition of convolution.
  14. Our perception of colour is based on three primaries, red, green and blue, but we do not usually describe a colour as red = 250, green 200 and blue = 100. Rather, we would say that a colour is a bright, pale orange and in this way we are describing a colour based on its intensity, hue and saturation. Although colour perception is subjective and particular to an individual, we can generally establish the relationship between RGB and IHS from the RGB colour cube model, which gives us elegant mathematical solutions of RGB–IHS and IHS–RGB transforms.
  15. The colour coordinate transformation based on the RGB colour cube assumes that our eyes respond equally to the three primary colours, when in fact they respond differently to each one. For the same intensity level, we generally feel that green is much brighter than blue and red. This is because our eyes are most sensitive to the intensity variation of green despite the fact that our brains receive strong stimulation from red and we can see red much farther away than the other two primaries. In a false colour composite of the same three bands, the visual effect may, however, be quite different. Any feature displayed in red appears far more obvious and distinctive than if it is displayed in green or blue. Sometimes, simply by changing the colour assignment of a colour composite image, you may visually enhance different ground objects.
  16. The saturation of a colour is generally regarded as the purity of a colour but the actual physical

- meaning of colour purity is rather vague and unclear; there are several different definitions of saturation as a result. According to the RGB colour cube model, saturation is not independent of intensity; it is the portion of white light in a colour. A large portion of white light in a colour produces low saturation. The concepts and calculation formulae of saturation in this book are given following this logic. Some older definitions of colour saturation ignore the dependency of saturation on intensity; they are not correct for the colour coordinate transformation based on the RGB colour cube.
17. We have introduced three image fusion techniques: intensity replacement (via the RGB–IHS transformation), the Brovey transform and SFIM. All of these can be used to improve the spatial resolution of colour composites but the RGB–IHS and Brovey transform techniques are subject to colour (spectral) distortion if the total spectral range of the low-resolution colour composite is different from that of the high-resolution image. SFIM is a spectral preservation data fusion technique which introduces no colour distortion, although its merits are somewhat offset by its demand for precise image co-registration, a problem which can be resolved by using the recently developed optical-flow-based, pixel-to-pixel image co-registration technique.
  18. Principal component analysis achieves the representation of a multi-spectral (or, more generally, multi-variable) image dataset in an orthogonal coordinate system defined by the axes of its data cluster ellipsoid. Mathematically, PCA performs coordinate rotation operations from the covariance matrix of the data cluster. A PC image is a linear combination of original image bands based on its corresponding eigenvector of the covariance matrix. Obviously, the data cluster of a multi-spectral image may not formulate a perfect ellipsoid and the PCA is therefore a statistical approximation.
  19. PCA produces  $n$  independent PC images from  $m$  ( $m \geq n$ ) correlated bands of a multi-spectral image. In most image processing software packages, the PCA function produces exactly the same number of PCs as the image bands rather than fewer. The point here is that the high-rank PCs contain little information. For the six reflective spectral bands of a TM/ETM+ image, the  $PC_6$  image contain nearly no information but random and striping noise. The first five PCs thus effectively represent the image information from all six reflective bands and, in this sense,  $n < m$ . On the other hand, dropping any an image band may considerably change the information.
  20. The idea of DS (Decorrelation Stretch) is to reduce the inter-band correlation of three bands for colour composition by stretching the data cluster to make it spherical rather than elongated. As a result, a colour composite with richer and more saturated colours but without hue distortion can be produced for visual analysis. DS can be achieved via either saturation stretch (IHSDS) or PC stretch (PCADS). We have proved that saturation stretch reduces inter-band correlation while PC stretch increases saturation. DDS (Direct Decorrelation Stretch) is the simplest and most effective technique for DS, which stretches saturation via a reduction of the achromatic component of colour without the involvement of forward and inverse transformations.
  21. Instead of performing the coordinate rotation from the covariance matrix, as PCA does, the image coordinate system of the original image bands can be rotated in any direction by a matrix operation. The key point is to make a rotation that enhances the image information of specific ground object properties. The tasseled cap transformation is one such coordinate rotation operation; it produces component images representing the brightness, greenness and wetness of a land surface.
  22. Classification is a very broad subject. This book covers only a small part of this subject, focused on multi-variable statistical classification. Both unsupervised and supervised multi-variable statistical classifications of multi-spectral images involve the feature space partition of image data clusters on the basis of particular dissimilarity measures (classifiers or decision rules); these do not involve the spatial relationships among neighbouring pixels. Multi-spectral images can be spatially segmented into patches of different classes according to

- particular properties and spatial continuity. This type of classification is referred to as image segmentation and it is not covered by this book. Hybrid multi-variable statistical classification in combination with image segmentation is also a valuable approach. The neural network is another branch of classification characterized by its self-learning ability, general applicability and great computational inefficiency.
23. The ultimate assessment of the accuracy of the image classification of ground objects must be based on ground truth, even though the ground truth data themselves can never be 100% accurate. Facing such a paradox and the practical difficulties of ground truth data collection, several image-data-based statistical methods have been proposed and, among them, the confusion matrix is the most popular. The confusion matrix method produces a user's accuracy and a producer's accuracy, as measures of the relative assessment of classification accuracy. These measurements are actually more applicable to the assessment of classification algorithm performance than to the indication of true classification accuracy.
24. Image co-registration and image rectification are closely related but different issues. Image rectification is used to rectify an image in compliance with a particular coordinate system. This can be done by co-registering the raw image to a map or to a geocoded image. In the second case, the rectification is achieved via image co-registration. Image co-registration is used to make two images match one another but not necessarily any map projection or coordinate system. Image warping based on a polynomial deformation model derived from ground control points (GCPs) is one of the most widely used techniques for co-registration between two images or between an image and a map. The technique is versatile in dealing with images that have significantly different scale and distortion but, since it is based on a deformation framework or mesh, it does not achieve accurate co-registration at every image pixel. A different approach is pixel-to-pixel co-registration based on accurate measurements of displacement between two images, at every pixel position. This can be achieved by local feature matching techniques, such as phase correlation. The technique ensures very high-quality co-registration but its application is limited by its low tolerance to differing acquisition conditions between the two images (e.g. scale, illumination angle, spatial features and physical properties).
25. We briefly introduced InSAR technology since it has become a very important tool in Earth observation remote sensing. InSAR data processing is a significant subject in its own right but it shares common ground with image processing, in optimizing the visualization of interferograms and analysing coherence images. For instance, ratio coherence is a typical image enhancement technique for separating spatial decoherence and temporal decoherence.

## 23.2 Geographical information systems

1. The fundamental building blocks of all GISs are the data structures – how we encode our information right from the start can have far-reaching consequences when we arrive at the more in-depth analytical stages of the work. The old adage of ‘rubbish in, rubbish out’, while seeming rather crude, can hold true. You must make sure that you choose the right data structure and that it is sensibly and intelligently structured for your needs. You must also ensure that it is captured as accurately as possible. At this stage, thinking about what you intend to do with the data will help when deciding how to structure them; try to think several steps ahead. From the basic units of point and pixel, everything else grows.
2. For raster data we must consider the need to describe adequately the spatial variability of a phenomenon, bearing in mind the effect of spatial resolution and data quantization on everything that we do beyond this step. On the other hand, we do not want to waste space and so we try to avoid redundancy wherever possible; these days, we need worry slightly less about this since increasingly efficient (almost lossless) data compression techniques are here to stay.

3. Consider very carefully whether you need to generate full topology in your vector dataset (remember the advice about thinking ahead of the analytical stages). If you do, you will need to do this from the start or face duplicating effort later on. On the other hand, you may not need it at all, depending on what you intend to do with your data, in the light of new, intelligent, vector data structures and fast computers.
4. You may not realize it at the start but there will probably be many instances when you need to convert data between raster and vector formats, and vice versa. Clearly there are implications for the methods we choose and kind of data involved. When you do make such conversions, it is important to pay attention to the detail at pixel level, so that you are fully aware of what happens to your data.
5. It is impossible to underestimate the importance of georeferencing – it is what holds your GIS together! If you are at all concerned about the accuracy and precision of your results, then take care to understand your coordinate space (datum and projection) and the effect that it has, regardless of the application, whether it be mapping, measuring or planning. Seemingly small differences on the screen could translate to sizeable errors on the ground. Learn how to use your equipment properly and understand the limitations of the device and what it tells you. Understanding the implications of getting it wrong should be a good incentive but it is better that you learn first!
6. Also, bear in mind the spatial resolution of any raster data you are working with and the effect this has on the precision and accuracy of the data you are recording or encoding. Also consider where in the world you are working and how this may affect your processing or calculations. For instance, consider what happens to the area represented by a square image pixel when you transform a raster dataset between polar and planar coordinate systems at high latitudes. The pixel will no longer be square and the data may need to be resampled, and may be resampled in the background without your being aware of it; the data values will be changed as a result. If the raster data have been generated by interpolation from point sample data, make sure the sample data are transformed from polar to planar coordinates before performing the interpolation, not after, to avoid this problem. Following on from this, it is advisable to ensure that all raster datasets input to your model conform to the same coordinate system and spatial resolution at the start; in fact, many software packages demand this.
7. Operations represent the greatest area of overlap between image processing and GIS. These are our principal tools for manipulating raster and vector data. The map algebra concept was born out of raster processing but can also be applied to vector data. In fact there are many parallel operations; their behaviour may be different, because of the differing data structures, but the principles and end results are the same. Sometimes the combination of two raster operations can be achieved using one vector operation; in such situations, it may be better to use the vector route to achieve the result but this will only be the case for categorical data. Once again, understanding the principles of the process at pixel (and point) level will help you to grasp the overall meaning, relevance and reliability of the result. The fundamental link with image processing in most of these processes should always be borne in mind; if in doubt, refer back to Chapter 14.
8. The use of the null (NoData) value is extremely important, so being aware of its uses and limitations will help to ensure that you achieve the right effect. Boolean logic is extremely useful in this and other contexts; it should be used wisely and carefully in conjunction with zeros and nulls.
9. Reclassification is probably one of the most useful tools in spatial analysis, for the preparation of both raster and vector data. Understanding the nature of the input data value range and type, for example via the histogram, is crucial. We stress the importance of considering the data type of input and output layers (whether integers or real numbers, for instance, as pointed out in Section 23.2, point 8), especially when using any arithmetic or statistical operations. After all, the output values should make reasonable sense.

10. We find that the geometric or attribute characteristics of one layer can be used to control operations performed on another, and so can be used to influence the path of the analysis in an indirect way, and that this can be done using raster data, vector data or both. Remember also that operations do not necessarily have to be arithmetic or involve calculations to change the attributes and/or geometric properties of features, as is the case with reclassifications and mathematical morphology.
11. Geostatistics is an extremely broad subject area which grew out of mineral exploration but now is applied in a great many fields. The key point of this topic is good data understanding: you need to understand the nature of the data at the start, i.e. through the use of histograms and semi-variograms, and through knowing how the data were collected and processed before you received them. It is far too easy to proceed without appreciating the characteristics of the dataset, such as the possibility of hidden trends in the data, the presence of multiple statistically independent populations, and/or the effects of spatial autocorrelation.
12. The next step is the choice of interpolation method and this can be mind boggling. The choice should be made according to the quantity, distribution and nature of your data, and on your intended use of the result, rather than on the ease of use of the tools at hand. In the end, you may not have a choice but you must understand these things first, if you are to make sense of and judge the quality of the result. Sometimes the most sophisticated method may be a waste of effort or may give meaningless results. On the other hand, using too simple a method may smooth out important detail or may not ‘honour’ the input data values. So it is important to consider your priorities and the intended use of the result, in choosing your approach.
13. The exploitation of surface data is another area of strong overlap with image processing, in image convolution filtering. In image processing, we use the neighbourhood process of filtering to enhance spatial image textures, for instance, whereas here we are trying to extract quantitative measures of surface morphology, measures which describe the 3D nature of that surface. Surfaces are extremely useful and powerful in a number of contexts; there is a lot more to it than merely visualizing them in three dimensions. Having said that, effective visualization can be achieved in two dimensions as well as three; GIS is effective in communicating information using either of these.
14. Do not forget that surfaces can be described by both raster and vector, so once again try to think about the intended use of the surface before you decide what kind of data you need to use. Surfaces are in themselves a rich data source from which several other valuable attributes (parameters) can be extracted. We give a selection of surface parameters that are particularly useful to geoscientists of all kinds, such as slope, aspect and curvature, describing how they are calculated as well as how they might be used. We stress the very direct link between the extraction of these surface parameters and their image processing equivalents, so if you need to understand their derivation from first principles more fully, then we refer you to Chapter 4.
15. There are also many potential sources of surface data and these are increasing in number and detail all the time. We provided a list of DEM data sources in Chapter 16 and some online resources, including the URLs of several useful sites for public data distribution, are listed in Appendix B.
16. Decision making represents another extremely broad subject area and one which extends well beyond digital mapping and the processing of spatial information; it is a kind of ‘cautionary tale’ which is relevant to all kinds of digital spatial analysis. The uncertainty we refer to affects everything we do in GIS and we need to be very aware of both the strengths and the limitations of our data, objectives and methods (and how they affect one another). This is the key to the successful application of the right kind of decision rules, to increasing the likelihood of making ‘good decisions’ and thus the reliability of the result, and to gaining the maximum reduction of ‘risk’ associated with a particular outcome.

17. Uncertainty creeps into our analysis in several different ways and in different forms, so we should try to understand the different situations in which these might occur, to help us handle the data correctly in the next stage of the analysis, for instance in constructing our decision rules and then in choosing the most appropriate multi-criteria combination method. There are several ways of dealing with uncertainty in spatial analysis; you may need to use some or all of them. Being aware of the potential problem of uncertainty and doing something about it is, in general, more important than the particular method you choose to deal with it.
18. Multi-criteria analysis is an extremely broad topic, so Chapter 18 could easily have been expanded into three or four separate and weighty chapters but we felt that this would be more detail than was necessary in this book. We have therefore tried to give an overview of the potential methods, listed broadly in increasing complexity, and describing their strengths and weaknesses, to enable you to select the best methods for your own case, since the choice is varied. Data preparation is still the most vital step, and one which will require the revision of all the preceding chapters, especially while you are considering the structure of your ‘model’ and selecting input evidence layers, the values they contain and the way they are scaled, before you proceed to the combination method.
19. The most appropriate method of layer combination for your situation will depend on the nature of the case, which will then dictate the set of decision rules; these will depend on the quantity and quality of the data and on your understanding of the problem. In complex problems that involve natural phenomena, it is highly likely that the input evidence may not all contribute equally to the outcome. Hence you may need to rank and weight the inputs or you may need to control the influence they have within the model you use. Something that might not be apparent at the start is the possibility that all the evidence may not contribute in the same way towards the problem; in other words, the absence of something may be as significant as the presence of something else.
20. Such sophisticated tools as are available today mean that a highly complex and impressive result is relatively easy to produce. Being able to repeat, justify and interpret that result is even more important, so some method of validation becomes increasingly important too. One last important thing to remember in this context is that there is no absolutely correct method for combining multiple criteria in spatial analysis but that, for any particular case, there is usually a most appropriate method; the world is indeed a fuzzy place.

### 23.3 Final remarks

There is clearly considerable common ground and synergy between the two technologies of image processing and GIS, and each one serves the other in some way. In our clear explanations in this book, we hope to have gone some way to demystifying the techniques which may previously have seemed inaccessible or a little daunting. These are all very useful tools which we have learned to use over the years and, from our collective experience and knowledge, into which we have also tried to inject some common sense and guidance wherever possible.

It is our intention that this book will serve as a key to the door of the active and integrated research fields of image processing and GIS and their applications within remote sensing, and this is why we have used the word ‘essential’ in the title of the book. We hope that the book is easy to read and to use, but how useful it will be depends entirely on the readers’ desire to explore.



# Appendix A

## Imaging Sensor Systems and Remote Sensing Satellites

The key element in remote sensing technology is the sensor. A sensor detects and quantitatively records the electromagnetic radiation (EMR) from an object remotely; hence the term ‘sensing’. For an object to be sensed, it must radiate energy either by reflecting the radiation impinging on it from an illumination source, or by ‘shining’ by itself. If a sensor provides its own illumination source it is an *active sensor* otherwise, if it depends on an independent illumination source, such as the Sun, or the radiation from the target itself, such as the Earth’s thermal emission, it is then a *passive sensor*. Synthetic aperture radar (SAR) is a typical active sensor system as it sends microwave radiation pulses to illuminate the target area and receives the returned signals to produce an image. In contrast, the most commonly used panchromatic and multi-spectral optical sensors are typical passive sensors. A camera can become an active sensor when it is used with flash light in the dark. In this case, the camera provides its own light source to illuminate the object and meanwhile takes a picture.

### A.1 Multi-spectral sensing

As a passive sensor, a multi-spectral imaging system images the Earth by recording either the

reflected solar radiation or the emitted radiation from the Earth. The Sun is the primary illumination source for the Earth. For Earth observation remote sensing, most passive sensor systems operate under solar illumination during the daytime; such systems range from aerial photography to satellite-borne multi-spectral scanners. These sensors detect *reflected* solar energy from the land surface to produce panchromatic and multi-spectral images. Features in these images are mainly described by two types of information: *spatial patterns* (e.g. topographic variation) and *spectral signatures*. Ignoring the minor factors, we can present such an image as

$$M_r(\lambda) = \rho(\lambda)E(\lambda) \quad (\text{A.1})$$

where  $M_r(\lambda)$  is the reflected solar radiation of spectral wavelength  $\lambda$  by the land surface, or an image of spectral band  $\lambda$ ,  $E(\lambda)$  is irradiance, that is the incident solar radiation energy upon the land surface, while  $\rho(\lambda)$  is the reflectance of the land surface at wavelength  $\lambda$ .

$E(\lambda)$  is effectively the topography, as determined by the geometry of the land surface in relation to illumination. The spectral reflectance,  $\rho(\lambda)$ , is a physical property that quantitatively describes the reflectivity of materials on the land surface at wavelength  $\lambda$ . The selective absorption and reflection by a material result in variation of spectral reflectance in a spectral range, giving a unique

signature for this substance. It is therefore possible to determine the land cover types or mineral compositions of the land surface based on spectral signatures using multi-spectral image data. Reflective spectral remote sensing is one of the most effective technologies for studying the Earth's surface as well as that of other planets.

The American Landsat satellite family, Thematic Mapper (TM) and Enhanced Thematic Mapper plus (ETM+), and the French SPOT satellite family HRV (High-Resolution Visible) are the most successful Earth Observation systems, providing broadband multi-spectral and panchromatic image data of global coverage. As shown in Table A.1, this type of sensor system operates in the visible spectral range with bands equivalent to three primary colours: blue (380–440 nm), green (440–600 nm) and red (600–750 nm); as well as in the near-infrared (NIR) (750–1100 nm) and shortwave-infrared (SWIR) (1550–2400 nm) ranges. The number of bands and the band spectral width in the VNIR (Visible Nearer Infrared) and SWIR spectral ranges are dictated by atmospheric windows and sensor design. For instance, the spectral width of the SWIR bands needs to be much wider than the visible bands if the same spatial resolution is to be achieved. This is the case for TM/ETM+ bands 5 and 7, because the solar radiation in the SWIR is significantly weaker than that in the visible spectral range.

In general, the term 'broadband' means that the spectral range is significantly wider than a few nanometres, as in the case of the hyperspectral sensor system described later. Broadband reflective multi-spectral sensor systems are a successful compromise between spatial resolution and spectral resolution. With relatively broad spectral bands, a sensor system offers reasonable spatial resolution with high SNR and, while operating in a wide spectral range from VNIR to SWIR, can provide images of multi-spectral bands enabling the identification of major ground objects and the discrimination of various land cover types. With dramatic improvements in sensor technology, from mechanical scanners to push-broom scanners, and to digital cameras, the spatial resolution of broadband multi-spectral imagery is improving all the time. For Sun-synchronous near-polar orbiting satellites, spatial resolution has been improved from 80 m (Landsat MSS) in the 1970s to a few metres and even sub-

metres on current systems, as shown by the examples in Table A.2.

The VNIR spectral range is used by nearly all the broadband reflective multi-spectral sensor systems. This spectral range is within the solar radiation peak and thus allows the generation of high-resolution and high-SNR images. It also covers diagnostic features of major ground objects, for instance:

- *Vegetation*: Minor reflection peak in green, absorption in red and then significant reflection peak in NIR. The phenomenon is often called 'red edge'.
- *Water*: Strong diffusion and penetration in blue and green and nearly complete absorption in NIR.
- *Iron oxide (red soils, gossans, etc.)*: Absorption in blue and high reflectance in red.

Many satellite sensor systems choose not to use the blue band, to avoid the very strong Rayleigh scattering effects of the atmosphere that make an image 'hazy'. A popular configuration is to offer three broad spectral bands in green, red and NIR, such as the case for SPOT, and the most recent high-spatial-resolution spaceborne sensors (Tables A.1 and A.2).

The SWIR spectral range is regarded as the most effective for lithological and mineral mapping because most rock types have high reflectance in the range 1.55–1.75  $\mu\text{m}$  and because hydrous (clay) minerals (often products of hydrothermal alteration) have diagnostic absorption features in the spectral range 2.0–2.4  $\mu\text{m}$ . These two SWIR spectral ranges correspond to Landsat TM/ETM+ bands 5 and 7, and to ASTER band 4 and bands 5–9 (Table A.1). SWIR sensor systems are technically more difficult and complicated because the SWIR detectors have to operate at very low temperatures, which therefore require a cooling system (a liquid nitrogen coolant or a cryo-cooler) to maintain the detectors at about 80 K.

With six broad reflective spectral bands, Landsat TM has provided the best spectral resolution of the broadband sensor systems for quite some time. The six broad reflective spectral bands are very effective for the discrimination of various ground objects but they are not adequate for the specific identification of rock types and mineral assemblages

**Table A.1** Comparison of the spectral bands of TM/ETM+, ASTER and SPOT

Sensor system	Terra-1 ASTER			Landsat-3–7 TM/ETM+			SPOT-1–3 HRV, SPOT-4 HRV1, SPOT-5 HRG		
	Band	Spectral range (μm)	Spatial res. (m)	Band	Spectral range (μm)	Spatial res. (m)	Band	Spectral Range (μm)	Spatial res. (m)
VNIR	1	0.52–0.60	15	1	0.45–0.53	30	1	0.50–0.59	20
	2	0.63–0.69		2	0.52–0.60		2	0.61–0.68	
3N	3N	0.78–0.86		3	0.63–0.69		3	0.79–0.89	
3B	3B	0.78–0.86		4	0.76–0.90		Pan	SPOT-1–3: 0.51–0.73 SPOT-4: 0.61–0.68 SPOT-5: 0.48–0.71	10 10 2.5–5
			ETM+ Pan	0.52–0.90	15				
SWIR	4	1.60–1.70	30	5	1.55–1.75	30	4	1.58–1.75	20
	5	2.145–2.185		7	2.08–2.35			Band 4 only on SPOT-4 HRV1 (High-Resolution Visible Infrared) and SPOT-5 HRG (High -Resolution Geometric)	
6	6	2.185–2.225							
7	7	2.235–2.285							
8	8	2.295–2.365							
9	9	2.360–2.430							
TIR	10	8.125–8.475	90	6	10.4–12.5	TM 120			
	11	8.475–8.825							
12	12	8.925–9.275							
13	13	10.25–10.95							
14	14	10.95–11.65					ETM+ 60		

**Table A.2** Some satellite-borne VHR (Very High-Resolution) broadband sensor systems. Hyp = Hyperspectral

Satellite	Launch time and status	Spatial resolution (m)			Spectral range (μm)		
		Pan	MS	Hyp.	Pan	MS	Hyp.
GeoEye1	Sept. 2008 In validation	0.41	1.64	—	0.45–0.80 0.51–0.58 0.655–0.69 0.78–0.92	0.45–0.51 0.51–0.58 0.655–0.69 0.78–0.92	—
WorldView-1	Sept. 2007 In operation	0.5	—	—	0.45–0.90	—	—
Ikonos 2	24 Sept. 1999 In operation	1	4	—	0.45–0.90 0.52–0.61 0.64–0.72 0.77–0.88	0.45–0.53 0.52–0.61 0.64–0.72 0.77–0.88	—
Quickbird	18 Oct. 2001 In operation	1	4	—	0.45–0.90 0.52–0.60 0.63–0.69 0.76–0.89	0.45–0.52 0.52–0.60 0.63–0.69 0.76–0.89	—
Orbview-3	26 June 2003 In operation	1	4	—	0.45–0.90 0.52–0.60 0.625–0.695 0.76–0.90	0.45–0.52 0.52–0.60 0.625–0.695 0.76–0.90	—
Orbview-4	Failed to orbit, Sept. 2001	1	4	8	0.45–0.90 0.52–0.60 0.625–0.695 0.76–0.90	0.45–0.52 0.52–0.60 0.625–0.695 0.76–0.90	0.45–2.50 200 bands

(of the hydrated type mentioned above, which are pathfinders for economic mineral deposits). Herein lies the demand for a sensor system with a much higher spectral resolution, at a bandwidth of a few nanometres, to detect the very subtle spectral signatures of materials on the land surface. This demand has led to the development of hyperspectral sensor systems.

ASTER (Advanced Spaceborne Thermal Emission and Reflection Radiometer), a push-broom scanner for VNIR and SWIR bands, represents a ‘transitional’ sensor system, somewhere between broadband multi-spectral and hyperspectral narrowband sensing. It is an integrated system of three scanners: a VNIR push-broom scanner with three broad spectral bands; an SWIR push-broom scanner with six narrow spectral bands; and a TIR (Thermal Infrared) across-track mechanical scanner with five thermal bands (Table A.1). The system

combines good spatial resolution in the VNIR bands with high spectral resolution in SWIR and multi-spectral thermal data which are very useful in geological applications. The three VNIR bands of 15 m resolution are adequate for distinguishing broad categories of land surface such as vegetation, water, soils, urban areas, superficial deposits and general rock outcrops while the six narrow SWIR bands of 30 m resolution have the potential to map major mineral assemblages (rock forming and alteration) and lithologies. Another unique advantage of ASTER is that it has along-track stereo capability. The VNIR scanner has a backward-viewing telescope to take NIR images in addition to its nadir telescope for the three VNIR bands. Thus nadir and backward-viewing NIR images are taken simultaneously, forming along-track stereo image pairs. These pairs enable generation of DEM (Digital Elevation Model) data.

Thus far, we have not specifically mentioned the panchromatic band image. We can regard panchromatic imagery as a special case of broadband reflective multi-spectral imagery with a wide spectral range, covering a large part of the VNIR spectral range, which can achieve a high spatial resolution.

## A.2 Broadband multi-spectral sensors

Aerial photography, using a large-format camera, is the earliest operational remote sensing technology for topographic surveying. Spaceborne remote sensing, in Earth Observation, began on Landsat-1 with its MSS (Multi-Spectral Scanner) and RBV (Return Beam Vidicon) camera, which was launched on 23 July 1972. These instruments captured and transmitted electronic images of the Earth; these images were then distributed to users in a digital format as digital image data for the first time. The concept and technology of digital imagery gave birth to satellite remote sensing. An Earth Observation satellite images the surface continuously from orbit and sends the images back electronically to the receiving stations on the ground.

The development of sensor technology is now mainly focused on improving spatial resolution and spectral resolution. For a given sensor system, its spatial resolution is dictated by the minimal energy level of electromagnetic radiation (EMR) that can make a signal distinguishable from the electronic background noise of the instrument, i.e. the dark current. This minimum energy of EMR is proportional to the product of radiation intensity over a spectral range, IFOV (Instant Field Of View) and the dwell time.

The IFOV is decided by the spatial sample density of an optical sensor system and it determines the pixel size of the image. For a given dwell time (equivalent to exposure time) and spectral range, the larger the IFOV, the more energy will be received by the sensor but the spatial resolution will be lower. To improve spatial resolution, the IFOV must be reduced, but to maintain the same energy level, either the dwell time or spectral range, or both, must be increased. When a sensor, which has a dwell time fixed by the sensor's design and platform orbit parameters, receives reflected

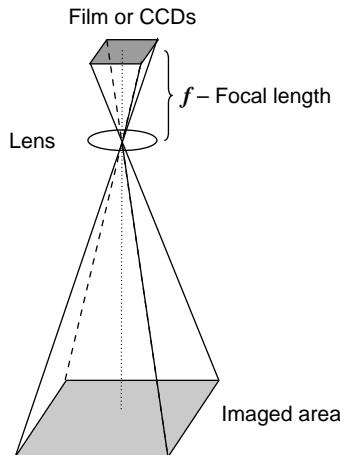
solar radiation from the Earth, it may record the energy in a broad spectral range as a single image, that is a panchromatic image, at a relatively high resolution. It may also split the light into several spectral bands and record them separately in several images of narrower spectral range, that is multi-spectral images. In this case, the energy that reaches the CCDs of each narrow spectral range is significantly weaker than in the panchromatic mode. To achieve the same energy level, the only solutions are to increase the size of the IFOV or to reduce spatial resolution. This is the reason that nearly all optical imaging systems achieve higher resolution in their panchromatic band than in the multi-spectral bands (Tables A.1 and A.2). For instance, the SPOT-1–3 HRV panchromatic band has 10 m resolution while XS (multi-spectral) bands have 20 m resolution.

The only other way to improve the spatial resolution in both panchromatic and multi-spectral imagery is to increase the dwell or exposure time. This has been an important consideration in sensor design, although the capacity for increasing the dwell time is very limited, for both airborne and spaceborne remote sensing, since the image is taken from a moving platform: long exposure time will blur the image.

### A.2.1 Digital camera

With few exceptions, passive sensor systems are essentially optical camera systems. As shown in Figure A.1, the incoming energy to the sensor goes through an optical lens and is focused onto the rear focal plane of the lens where the energy is recorded by radiation-sensitive media or a sensor device, such as film or CCD (Charged Coupled Device).

A digital camera is built on a full 2D CCD panel; it records an image through a 2D CCD panel linking to a memory chip. With the advances in multi-spectral 2D CCD technology, the new generation of passive sensors will be largely based on the digital camera mechanism that takes an image in an instantaneous frame rather than scanning line by line. The consequence is that the constraints on platform flight parameters can be relaxed, image resolution (spatial and spectral) can be improved and image geometric correction processing can be streamlined.



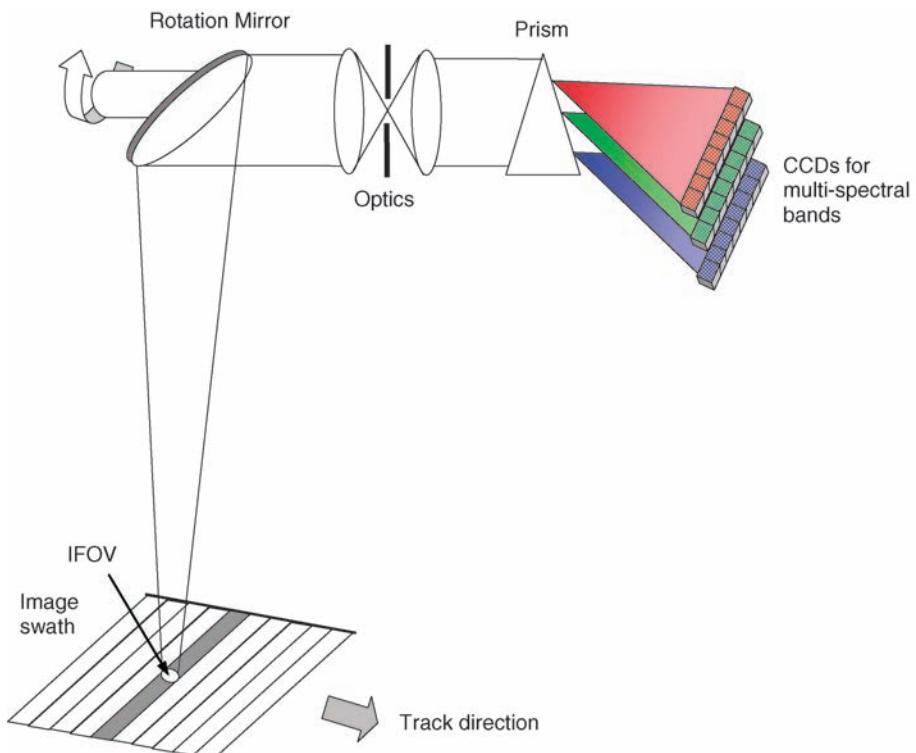
**Figure A.1** Basic structure of an optical sensor system

### A.2.2 Across-track mechanical scanner

The early design of spaceborne passive sensor systems was constrained by the technology of the primary sensor unit, the CCD. The mechanical scanner has been the ideal solution to achieve multi-spectral imaging at relatively high resolution, since it is based on a simple mechanical device which uses only a few CCDs. Figure A.2 contains a schematic diagram showing the principle of an across-track mechanical multi-spectral scanner. The optical part of a scanner is essentially a camera but the image is formed pixel by pixel, scanned by a rotating mirror, and line by line, as the platform (aircraft or satellite) passes over an area. The range of a scan line is called the *swath*. The light reflected from the land surface reaches the rotating mirror that rotates at a designated speed and thus views different positions on the ground along a swath during its rotation scan cycle. The rotating mirror diverts the incident light through the scanner optics and then the light is dispersed into several spectral beams by a spectral splitting device (a prism or interference filters). The multi-spectral spectral beams are then received by a group of CCDs which sample the light at a regular time interval. By the time the scanner finishes scanning one swath, the satellite or aircraft has moved forward along its track to the position for the next scan. One scan swath can be a single image

line or several image lines depending on the sensor design and the synchronization between flying speed, swath width, altitude of the satellite or aircraft and required image resolution. In this way, a scanner can achieve a very large image using limited CCDs; although mechanically rather complex, a scanner relies less on the CCD technology.

The MSS, with four spectral bands onboard Landsat-1–3, is a classical example of a mechanical scanner. It is a one-way scanner that scans in one direction of mirror rotation only, and with an empty return run. Such a design makes compensation for the Earth's rotation easier, since the Earth rotates a fixed distance along the swath direction in each scanning cycle. The inactive return runs waste valuable time for imaging and cause a shorter dwell time in the active runs, thus reducing image spatial resolution. The TM/ETM+, onboard Landsat-4–7, is a significantly improved scanner with six reflective spectral bands and one thermal band (Table A.1). It is a two-way scanner that scans in both directions. So, for the same width of swath, the two-way scan allows the mirror to rotate more slowly, thus increasing the dwell time of the CCDs at each sample position. This configuration improves both spatial and spectral resolution. To compensate for the Earth's rotational effects, the geometric correction for TM/ETM+ is more complicated than that for MSS because one scan direction is *for*, and the other *against*, the Earth's rotation.



**Figure A.2** A schematic illustration of an across-track mechanical multi-spectral scanner

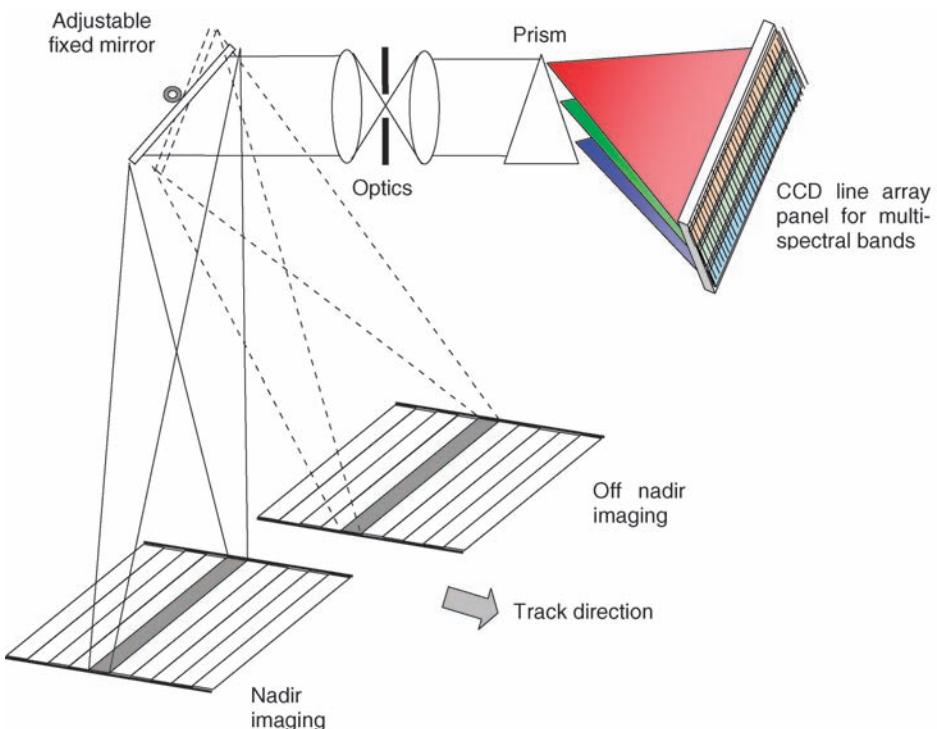
### A.2.3 Along-track push-broom scanner

With rapid developments in CCD technology, a more advanced push-broom scanner has become dominant in broadband multi-spectral sensor design since the successful launch of SPOT-1 on 22 February 1986. As shown in Figure A.3, the key difference between a push-broom scanner and a mechanical scanner is that it does not have a mechanical part for pixel-by-pixel scanning along the swath direction. Instead of a rotating mirror, a push-broom scanner has a steerable fixed mirror to enable the sensor to image its swath either at nadir or off nadir. A line array panel of CCDs covering the whole imaging swath is mounted at the rear of the spectral dispersion device. The push-broom scanner images an area, line by line, along the track when the sensor platform (a satellite or an aircraft) passes over, just like pushing a broom forward to sweep the floor.

Since one swath of image is generated simultaneously, the dwell time for each CCD representing an image pixel can be as long as a whole swath

scanning time for a mechanical scanner. With significantly increased dwell time, a push-broom scanner achieves much higher resolution. Based on advanced CCD technology, the push-broom scanner is also much simpler than the mechanical scanner in structure, and the data geometric correction is less complicated. With no mechanical parts, the system is robust and reliable. The number of CCDs in the line array, in the swath direction, decides the size and the resolution of the image data generated. For instance, the SPOT HRV has 6000 CCDs per line for its panchromatic band and 3000 CCDs per line for its multi-spectral bands. The system therefore produces panchromatic images of 6000 pixels per line at 10 m resolution and multi-spectral images of 3000 pixels per line at 20 m resolution.

The push-broom design also allows greater flexibility in manoeuvring the sensor to point at off-nadir positions. As shown in Figure A.3, unlike a mechanical scanner, the mirror fixed in front of the lens is not for scanning but to direct



**Figure A.3** A schematic diagram of an along-track push-broom multi-spectral scanner

the incident light to the optics. It is steerable with several fixed angle positions. At different angles, the mirror directs the sensor system to view and image different scenes, off nadir, along the satellite's track.

### A.3 Thermal sensing and thermal infrared sensors

A thermal sensor is also passive but the radiation energy that the sensor receives has been emitted from the Earth's surface rather than reflected from it. Thermal sensing does not therefore need an illumination source since the target itself is the illumination source. The Earth's surface can be approximated as a black body of 300 K and, using Wien's law, we can calculate that the radiation peak of the Earth is at about 10  $\mu\text{m}$ . In this spectral range, radiation can be sensed and measured by temperature rather than visible brightness; it can therefore be called thermal sensing. Different natural materials on the land surface have different thermal radiation properties

and thermal sensors are therefore useful tools for geological and environmental studies.

There are quite a few airborne TIR sensor systems, for example the thermal infrared multispectral scanner (TIMS), with six bands in the 8.2–12.2  $\mu\text{m}$  spectral region, was developed in 1982. Landsat TM and ETM+ have a broad thermal band at a wavelength of 10.4–12.5  $\mu\text{m}$ . ASTER onboard the Terra-1 satellite comprises a multi-spectral thermal system with five narrow thermal bands as listed in Table A.1.

In general, a broadband TIR sensor operating in the 8–14  $\mu\text{m}$  spectral range images the radiation temperature of the land surface while the narrower band multi-spectral thermal image data present the thermal spectral signatures of materials of the land surface. It is important to know that daytime thermal images are fundamentally different from those acquired at night. Daytime thermal images are dominated by topography, as governed by the geometry between slopes and solar radiation, in the same way as in reflective multi-spectral images, whereas the pre-dawn night thermal images, which

are nearly solely determined by emission from the Earth's surface, show better the thermal properties of ground materials.

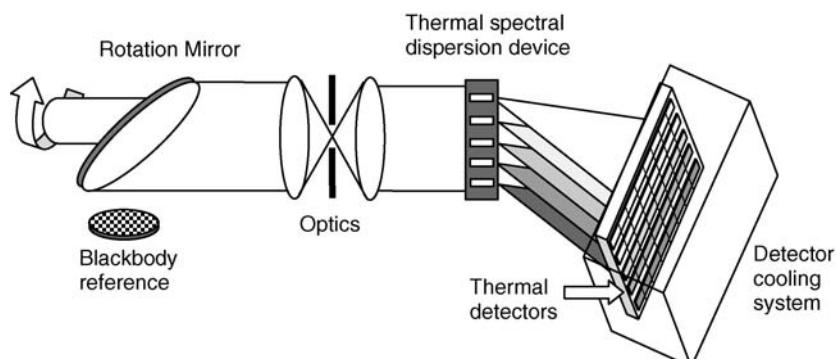
In both systems, TM/ETM+ and ASTER, the spatial resolution of thermal bands is significantly lower than that of the reflective multi-spectral bands, as shown in Table A.1. One reason is that the interaction between thermal energy (or heat) and the atmosphere is more complex than in the case of VNIR and SWIR energy. Heat can be transmitted in the air not only by radiation but also by air circulation. Secondly, the solar radiation impinging on the Earth in the TIR spectral range and the direct thermal emission from the Earth are both very weak compared with the energy intensity of the Earth's reflected solar radiation in the VNIR and SWIR spectral ranges.

So far, most thermal sensors are of the across-track mechanical scanner type as shown in Figure A.4. The major difference of a thermal scanner from a reflective multi-spectral scanner is that it needs a cooling system to maintain the TIR detector at very low temperature for maximum sensitivity. For instance, the thermal sensor of the Landsat TM is surrounded by liquid nitrogen at 77 K and stored in an insulated vessel. In the ASTER system, a cryo-cooler is used to maintain the TIR detectors at 80 K. A black body plate is used as an onboard calibration reference that is viewed before and after each scan cycle, thus providing an estimate of instrument drift. This is essential for maintaining the accuracy and consistency of a TIR instrument. The temperature sensitivity of a modern TIR sensor system can be as high

as 0.1 K. To represent the sensitivity fully, many thermal IR sensors use 10–12 bit quantization to record data, such as ASTER multi-spectral thermal band images which are 12 bit integer data.

## A.4 Hyperspectral sensors (imaging spectrometers)

Passive sensor technological development is continually aiming at higher spatial and spectral resolutions. Hyperspectral sensor systems represent a revolutionary development in the progress of optical sensor spectral resolution, which may be as high as a few nanometres, and can generate nearly continuous spectral profiles of land surface materials. A hyperspectral sensor system is a combination of the spatial imaging capacity of an imaging system with the spectral analytical capabilities of a spectrometer. Such a sensor system may have several hundred narrow spectral bands and a spectral resolution of the order of 10 nm or narrower. Imaging spectrometers produce a complete spectrum for every pixel in the image; the dataset is truly a 3D data cube which allows identification of materials, rather than mere discrimination as with broadband sensor systems. The data processing methodology and strategy are therefore different in many aspects from broadband images. It is more important to analyse the spectral signature for each pixel rather than to enhance the image to improve visualization, although the latter will still be essential later on.



**Figure A.4** A schematic diagram of a thermal scanner

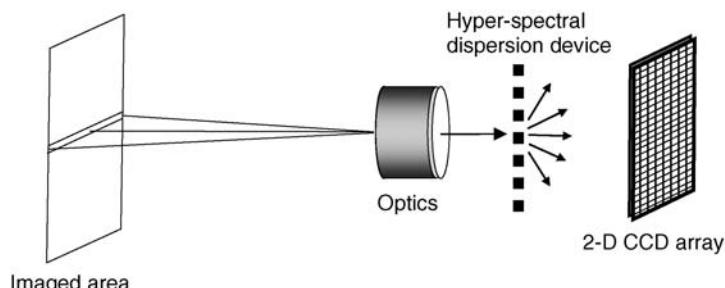
**Table A.3** Some hyperspectral sensors (airborne and satellite borne\*)

Instrument	Spectral range (nm)	Bandwidth (nm)	No. of bands
AVIRIS	400–2400	9.6	224
AIS	1200–2400	9.6	128
SISEX	400–2500	10–20	128
HIRIS	400–2500	10–20	128
MIVIS	430–1270	20–54	102
HyMAP	400–2500	16	125
Hyperion*	400–2500	10	242

One of the earliest and the most representative hyperspectral systems is JPL's advanced visible infrared image spectrometer (AVIRIS) (see Table A.3). Figure A.5 shows the general principle of hyperspectral systems. The incoming EMR from the land surface goes through the sensor optics and is then split into hundreds (e.g. 224 for AVIRIS) of very narrow spectral beams by a spectral dispersion device (e.g. interference filters) and finally the spectral beams are detected by arrays of CCDs corresponding to, for instance 224 spectral bands. A hyperspectral system can be either an across-track mechanical scanner, with a small number of detectors for each band, or an along-track push-broom scanner, with a panel of hundreds of line arrays of CCDs. Hyperspectral sensors are so far operating in either VNIR only or VNIR and SWIR spectral ranges (Table A.3).

As another type of passive sensor, microwave radiometers are designed to image the emitted radiation from the Earth's surface at this spectral range.

Thermal radiation from natural surfaces, such as the land surface, extends from its peak in the thermal infrared region into the microwave region. An Earth observation microwave imaging radiometer operates in this spectral region to receive microwave radiation from the Earth. As a passive sensor system, it is important to understand that a microwave radiometer is fundamentally different from a radar sensor which is a *ranging system*. The only similarity between the two is that they both operate in the microwave spectral range. A passive microwave sensor system, a microwave imaging radiometer, works more like a thermal sensor system. It collects emitted energy radiated from the Earth in the microwave spectral range and provides useful information relating to surface temperature, roughness and material dielectric properties. This type of sensor has been used for global temperature mapping, polar ice mapping and regional soil moisture monitoring.

**Figure A.5** Principle of an imaging spectrometer (hyperspectral system)

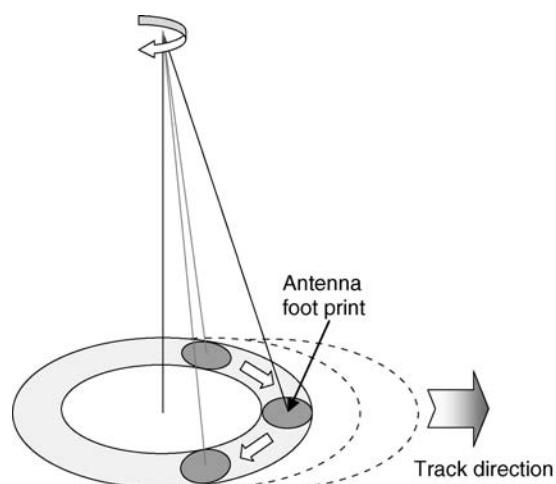
A spaceborne microwave imaging radiometer is often a multi-channel scanner such as the scanning multi-channel microwave radiometer (SMMR), onboard Seasat and Nimbus (1978), and the microwave imager, onboard the TRMM Tropical Rainfall Measuring Mission (TRMM) satellite (1997). It consists of an antenna together with its scanning mechanism, a receiver and a data handling system. The received emitted microwave signals are closely related to the observation angle and the path length in the atmosphere. Ensuring that these scanning parameters are constant can significantly increase the accuracy of the derivation of the surface parameters from microwave brightness temperature. A conical scan configuration is popular for passive microwave scanners. As shown in Figure A.6, the antenna observation direction is offset at a fixed angle from nadir, rotating its scan around the vertical (nadir) axis and thus sweeping the surface of a cone. If the scan is configured for full  $360^\circ$ , double coverage fore and aft of the spacecraft is obtained. With the forward motion of the satellite along its orbit, a belt of land surface is imaged. Obviously, in a conical scanning geometry, the observation angle and distance to any scanned position are constants. Spaceborne passive microwave scanners are usually of low spatial resolution, from several kilometres to several tens of kilometres, because of the weak signal in the microwave spectral range.

## A.6 Active sensing: SAR imaging systems

Radar is essentially a ranging or distance-measuring device. Nearly all the imaging radar systems are configured as *side-looking*, referred to as side-looking radar (SLR). The reason is that, as a ranging system, radar forms an image by recording the position of return signals based on time. If a radar system is configured to view both sides of the platform (aircraft and satellite) symmetrically, the return signals from both sides in an equal distance will be received at the same time, causing ambiguity.

A radar system transmits microwave pulses at a carrier wavelength/frequency and then receives the echoes of these pulses scattered back by ground surface objects. The wavelength and frequency (radar bands) of commonly used microwave pulse carriers are listed in Table A.4. The code letters for the radar bands in the table were allocated during the Second World War, and remain to this day.

Radar image data are configured in relation to two coordinates: slant range and azimuth. Slant range corresponds to the two-way signal delay time. By measuring the time delay between the transmission of a pulse and the reception of the backscattered ‘echo’ from different targets, their distance to the radar and thus their location can be determined, and, in this way, a radar image is built in the slant



**Figure A.6** The conical scanning mechanism of a passive microwave sensor

**Table A.4** Radar bands, wavelengths and frequencies

Band	Wavelength $\lambda$ (cm)	Frequency (MHz; $10^6$ cycles/s)
Ka	0.75–1.1	40 000–26 500
K	1.1–1.67	26 500–18 000
Ku	1.67–2.4	18 000–12 500
X	2.4–3.75	12 500–8 000
C	3.75–7.5	8 000–4 000
S	7.5–15	4 000–2 000
L	15–30	2 000–1 000
P	30–100	1 000–300

range. In the azimuth direction, the image is built according to the pulse number sequence. As the radar platform moves forward, it transmits microwave pulse beams to scan on one side of its flight path, strip by strip, and simultaneously records the backscattered signals. As such, a 2D radar image is built up.

The azimuth resolution  $R_a$  of a *real aperture* SLR is a function of radar wavelength  $\lambda$ , the slant range  $S$  and the radar antenna length,  $D_r$ :

$$R_a = \frac{S\lambda}{D_r}. \quad (\text{A.2})$$

According to this formula, the azimuth resolution  $R_a$  is inversely proportional to the length of the radar antenna  $D_r$ . For a given radar wavelength and slant range, the longer the antenna, the higher the azimuth resolution. There is, however, a physical limit to the length of a radar antenna onboard an aircraft or satellite, and that constrains the potential spatial resolution.

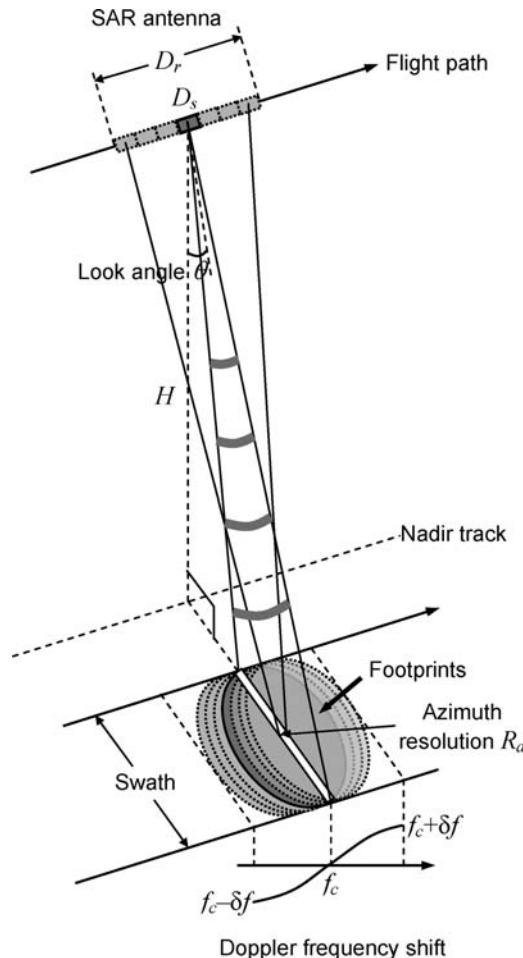
SAR is a technology which solves this problem. Compared with conventional, real aperture radar, SAR achieves high along-track (azimuth) resolution by synthesizing a virtual long antenna with the motion of a very short antenna, by intensive data processing of the coherent backscattered radar signals based on information of the Doppler frequency shift.

As illustrated in Figure A.7, while the SAR platform is moving along its path at an altitude  $H$ , it transmits microwave pulses into the antenna's illumination footprint, at the rate of the pulse

repetition frequency (PRF), and receives the echoes of each pulse backscattered from the target surface. The short antenna of an SAR produces a wide footprint (the area illuminated by the radar beam pulse) on the ground. If the platform motion is small in comparison with the width of footprint and the PRF, several consecutive footprints overlap. Typical PRFs for SAR systems are in the range of 1–10 kHz, which is relative high in relation to the low travel speed of a platform, and thus each point on the ground can be illuminated many times when an SAR platform passes over. The echoes of these repeat illuminations of the same point will be received by the antenna at a slightly higher frequency than the carrier frequency as  $(f_c + \delta f)$ , when the SAR is approaching the point, and slightly lower frequency than the carrier frequency as  $(f_c - \delta f)$ , when it is moving away from the point, according to Doppler frequency shift effects. Thus depending on its position in the overlapped stack of footprints, a point is uniquely coded by its Doppler frequency shift signature. Consequently, the signal processing of matched filtering from the Doppler effect can achieve very high azimuth resolution on the scale of a fraction of the SAR footprint width. The effect is equivalent to connecting a sequence of positions for a short antenna, of a SAR travelling along its path, corresponding to the overlapped footprints over a ground point, to formulate a very long virtual antenna as though it were a very long real aperture antenna to focus on the point. Obviously, a short SAR antenna means a wide footprint and thus allows more footprints to overlap, forming a long virtual antenna to achieve high azimuth resolution. It can be proved that the azimuth resolution of the SAR is half the length of its antenna (Curlander and McDonough, 1991), or the shorter antenna diameter  $D_s$  of SAR achieves higher azimuth resolution:

$$R_a = \frac{D_s}{2}. \quad (\text{A.3})$$

For a high-slant-range resolution, the SAR emits a chirp pulse with a bandwidth  $B_\nu$  of tens of megahertz modulating a carrier wave frequency  $f_c$  (the nominal frequency of the SAR). Depending on the increase or decrease in chirp frequencies, there are ascending or descending chirps. For the case of an ascending chirp, the exact frequency of a radar



**Figure A.7** Principle of SAR. The motion of the SAR antenna with a small length of  $D_s$  along its flight path simulates a virtual long antenna  $D_r$  that enables a high azimuth resolution  $R_a$  much smaller than the SAR footprint width, via matched filtering processing on the overlapped footprints (based on Doppler frequency shift  $\delta f$  from the SAR carrier frequency  $f_c$ )

pulse reaching the ground is higher in the far ground range than in the near ground range, and so are the echoes from different ground ranges. The same applies to a descending chirp. The returned signal is then demodulated with the chirp form and sampled based on the chirp frequency shift from the nominal frequency of the SAR, via matched filtering, to achieve high range resolution.

When discussing microwave energy, the *polarization* of the radiation is also important. Polarization refers to the orientation of the electronic field. Most radar systems are designed to transmit

microwave radiation either horizontally polarized (H) or vertically polarized (V). Similarly, the antenna receives either the horizontally or vertically polarized backscattered energy, and some radar systems can receive both. These two polarization states are designated by the letters H for horizontal, and V for vertical. Thus, there can be four combinations of both transmit and receive polarizations as follows:

HH – for horizontal transmission and horizontal receipt,

**Table A.5** Some past, present and future spaceborne SAR systems

Sensor	Country	Mission period	Band; polarization	Look angle (°)	Antenna size (m)	Alt. (km)	Swath (km)
Seasat-A	USA	June 1978, 105 d	L; HH	20	10.8 × 2.2	795	100
SIR-A	USA	Nov. 1981, 2.5 d	L; HH	47	9.4 × 2.2	260	50
SIR-B	USA	Oct. 1984, 8.3 d	L; HH	15–60	10.8 × 2.2	224, 257,	20–40
SIR-C/ X-SAR	USA, Germany, Italy Russia EU	Apr. 1994, 11 d	L, C, X; multi-pol.	20–55	12 × 2.9/ 12 × 0.7/12 × 0.4	360 225	360 15–90, 225
ALMAZ-1		Mar. 1991, 2.5 years 17 July 1991 to Mar. 2000	S; HH C; VV	20–65 20.335	12 × 1.5 10 × 1	300–370	30–45 100
ERS-2	EU	21 Apr. 1995 to present	C; VV	20.355	10 × 1	780	100
JERS-1	Japan	Feb. 1992 to Oct. 1998	L; HH	38	12 × 2.4	570	75
RADARSAT-1	Canada	4 Nov. 1995 to present	C; HH	20–60	15 × 1.5	790–820	50–500
SRTM	USA, Germany EU	11 Feb. 2000, 11 d	C; X; HH, VV	20–60		233	56–225
ENVISAT		1 Mar. 2002 to present	C; HH, VV, VH/HV	15–45	10 × 1	800	57–400
RADARSAT-2	Canada	14 Dec. 2007 to present	C; multi-pol.	Variable	15 × 1.5	790–820	20–500
PALSAR	Japan	24 Jan. 2006	L; multi-pol.	8–60	8.9 × 2.9	692	30–350
TerraSAR-X	Germany/UK	15 June 2007	X; multi-pol.	20–55	4.8 × 0.7	514	5 × 10 to 100 × 150
TerraSAR-2	Germany/UK	2010					
TerraSAR-3	Germany/UK	2015					

VV – for vertical transmission and vertical receipt,  
HV – for horizontal transmission and vertical receipt, and

VH – for vertical transmission and horizontal receipt.

The first two polarization combinations are referred to as *like polarized* because the transmitted and received polarizations are the same. The last

two combinations are referred to as *cross-polarized* because the transmitted and received polarizations are orthogonal.

Some past, present and future spaceborne SAR systems are listed in Table A.5. SAR image data are supplied in several different formats. Typically, Single Look Complex (SLC) data are in 8-byte complex numbers, while multi-look (intensity) images are in 16-bit unsigned integers.



# Appendix B

## Online Resources for Information, Software and Data

Here we have compiled a list of what we hope are useful resources available on or via the Internet (as of November 2008). These include links to the sites of proprietary software suites and to those providing

programs which are shareware, or low cost, or entirely free of licence. These sites are in themselves often rich sources of background information and technical help. Secondly, we include links to online

### B.1 Software – proprietary, low cost and free (shareware)

Autodesk	<a href="http://www.usa.autodesk.com">www.usa.autodesk.com</a>
ERDAS	<a href="http://www.erdas.com">www.erdas.com</a>
ER Mapper	<a href="http://www.ermapper.com">www.ermapper.com</a>
ESRI	<a href="http://www.esri.com">www.esri.com</a>
FME Safe Software	<a href="http://www.safe.com">www.safe.com</a>
Geotools	<a href="http://www.geotools.codehaus.org">www.geotools.codehaus.org</a>
GlobalMapper	<a href="http://www.globalmapper.com">www.globalmapper.com</a>
GRASS	<a href="http://www.grass.itc.it">www.grass.itc.it</a>
Idrisi	<a href="http://www.clarklabs.org">www.clarklabs.org</a>
ILWIS	<a href="http://www.itc.nl/ilwis">www.itc.nl/ilwis</a>
ITTVIS ENVI	<a href="http://www.ittvvis.com/envi">www.ittvvis.com/envi</a>
JUMP GIS	<a href="http://www.jump-project.org">www.jump-project.org</a>
Landserf	<a href="http://www.landserf.org">www.landserf.org</a>
Map Window	<a href="http://www.mapwindow.org">www.mapwindow.org</a>
MapInfo	<a href="http://www.mapinfo.com">www.mapinfo.com</a>
PCI Geomatics	<a href="http://www.pcigeomatics.com">www.pcigeomatics.com</a>
Quantum GIS opensource	<a href="http://www.qgis.org">www.qgis.org</a>
SAGA GIS	<a href="http://www.saga-gis.org">www.saga-gis.org</a>
Various independent	<a href="http://www.rockware.com">www.rockware.com</a>
Variowin	<a href="http://www.sst.unil.ch/research/variowin">www.sst.unil.ch/research/variowin</a>
Virtuozo	<a href="http://www.supresoft.com.cn/english/products/virtuozo/virtuozo.htm">www.supresoft.com.cn/english/products/virtuozo/virtuozo.htm</a>

information and technical resources which are largely independent of any allegiance to particular software or are provided by independent (often charitable) organizations. Thirdly, we include a list of online

data sources, some of which allow downloading of data (either free of charge or with some payment system in place) and others which merely enable browsing of available data as quick looks or listings.

## B.2 Information and technical information on standards, best practice, formats, techniques and various publications

Association for Geographic Information (AGI)	<a href="http://www.agi.org.uk">www.agi.org.uk</a>
British Geological Survey (BGS)	<a href="http://www.bgs.ac.uk">www.bgs.ac.uk</a>
Committee on Earth Observation Satellites (CEOS)	<a href="http://www.coes.cnes.fr">www.coes.cnes.fr</a>
Digital Earth	<a href="http://www.dgeo.org">www.dgeo.org</a>
Digital National Framework	<a href="http://www.dnf.org">www.dnf.org</a>
ESRI ArcUser online	<a href="http://www.esri.com/news/arcuser">www.esri.com/news/arcuser</a>
ESRI online knowledge base	<a href="http://www.support.esri.com/index.cfm?fa=knowledgebase.gateway">www.support.esri.com/index.cfm?fa=knowledgebase.gateway</a>
Geospatial Analysis online	<a href="http://www.spatialanalysisonline.com">www.spatialanalysisonline.com</a>
Geospatial Information and Technology Association (GITA)	<a href="http://www.gita.org">www.gita.org</a>
GIS Day	<a href="http://www.gisday.com">www.gisday.com</a>
GIS Research UK (GISRUK)	<a href="http://www.geo.ed.ac.uk/gisruk">www.geo.ed.ac.uk/gisruk</a>
Grid Forum 2001	<a href="http://www.gridforum.org">www.gridforum.org</a>
International Association of Photogrammetry & Remote Sensing	<a href="http://www.cse.dmu.ac.uk/EAPRS/IAG">www.cse.dmu.ac.uk/EAPRS/IAG</a>
International DEM Service	<a href="http://www.casa.ucl.ac.uk/software/isovist.asp">www.casa.ucl.ac.uk/software/isovist.asp</a>
Isovist Analyst	<a href="http://www.census.gov/geo/www/tiger/index.html">www.census.gov/geo/www/tiger/index.html</a>
MAF/TIGER background documents	<a href="http://www.opengeospatial.org">www.opengeospatial.org</a>
Open Geospatial Consortium	<a href="http://www.ordnancesurvey.co.uk">www.ordnancesurvey.co.uk</a>
Ordnance Survey (OS)	<a href="http://www.rspsoc.org">www.rspsoc.org</a>
Remote Sensing and Photogrammetry Society (RSPSoc)	<a href="http://www.ukgeoforum.org.uk">www.ukgeoforum.org.uk</a>
UKGeoForum (umbrella organization)	<a href="http://www.geovrml.org">www.geovrml.org</a>
Web 3D Consortium	<a href="http://www.w3.org">www.w3.org</a>
World Wide Web Consortium	

## B.3 Data sources including online satellite imagery from major suppliers, DEM data plus GIS maps and data of all kinds

ALOS data search	<a href="http://www.cross.restec.or.jp">www.cross.restec.or.jp</a>
Asia Pacific Natural Hazards Network	<a href="http://www.pdc.org/mde/explorer.jsp">www.pdc.org/mde/explorer.jsp</a>
Digital Globe (Quickbird & WorldView)	<a href="http://www/browse.digitalglobe.com">www/browse.digitalglobe.com</a>
EarthExplorer	<a href="http://www.earthexplorer.usgs.gov">www.earthexplorer.usgs.gov</a>

EOS DataGateway	<a href="http://edcimswww.cr.usgs.gov/pub/imswelcome/plain.html">edcimswww.cr.usgs.gov/pub/imswelcome/plain.html</a>
ESA EOLI catalogues	<a href="http://www.catalogues.eoportal.org">www.catalogues.eoportal.org</a>
GeoCommunity GIS free data depot	<a href="http://www.data.geocomm.com">www.data.geocomm.com</a>
GeoEye (GeoFUSE)	<a href="http://www.geofuse.geoeye.com/landing">www.geofuse.geoeye.com/landing</a>
GIS data depot	<a href="http://www.gisdepot.com">www.gisdepot.com</a>
GIS Lounge	<a href="http://www.gislounge.com">www.gislounge.com</a>
GLCF	<a href="http://www.lcf.umbc.edu">www.lcf.umbc.edu</a>
Glovis	<a href="http://www.glovis.usgs.gov">www.glovis.usgs.gov</a>
SPOT catalogue	<a href="http://www.sirius.spotimage.fr">www.sirius.spotimage.fr</a>
SRTM Public Data	<a href="http://www2.jpl.nasa.gov/srtm/cbanddataproducts.html">www2.jpl.nasa.gov/srtm/cbanddataproducts.html</a>



# References

In this book, we have organized the references to ensure optimum readership.

The references are listed in three parts: general references; Part One references and further reading; and Part Two references and further reading. The references relating to the case studies in Part Three are given at the end of each of the chapters.

The general references are divided into three sections, listing some of the most useful referenced books on image processing, GIS and remote sensing. These books cover much of the common ground with each other and with this book. Many of the widely used techniques and algorithms described in this book are derived from these books in conjunction with the authors' personal experience, judgement and modifications. For this reason, the general references are not generally directly cited in this book, since they can be relevant to any part of the book.

The publications listed in the references and further reading sections of Parts One and Two are referred to in the text if they are directly relevant to particular algorithms, techniques or applications. Publications which are of general relevance to one or several parts of the book are not directly referred to in the text but are cross-linked to the related contents via the index. All publications are directly referenced, or are indexed, or both.

## General references

### Image processing

- Castleman, K.R. (1996) *Digital Image Processing*, 2nd edn, Prentice Hall International, Englewood Cliffs, NJ.
- Gonzalez, R.C. and Woods, R.E. (2002) *Digital Image Processing*, 2nd edn, Pearson Education, Singapore.
- Gonzalez, R.C., Woods, R.E. and Eddins, S.L. (2004) *Digital Image Processing using MATLAB*, 1st edn, Pearson Prentice Hall, Englewood Cliffs, NJ.

Jensen, J.R. (2005) *Introductory Digital Image Processing – A Remote Sensing Perspective*, 3rd edn, Prentice Hall, Upper Saddle River, NJ.

Mather, P.M. (2004) *Computer Processing of Remotely-Sensed Images: An Introduction*, 3rd edn, John Wiley & Sons, Ltd, Chichester.

Niblack, W. (1986) *An Introduction to Digital Image Processing*, Prentice Hall International, Englewood Cliffs, NJ.

Richards, J.A. (2006) *Remote Sensing Digital Image Analysis: An Introduction*, 4th edn, Springer-Verlag, Berlin.

Schowengerdt, R.A. (1997) *Remote Sensing: Models and Methods for Image Processing*, 2nd edn, Academic Press, London.

## GIS

- Bonham-Carter, G.F. (2002) *Geographic Information Systems for Geoscientists: Modelling with GIS*, Pergamon/Elsevier, Amsterdam.
- Burrough, P.A. (1998) *Principles of Geographic Information Systems*, 2nd rev. edn, Clarendon Press, Oxford.
- Burrough, P.A. and McDonnell, R.A. (1998) *Principles of Geographical Information Systems*, Oxford University Press, New York.
- Demers, M.N. (2009) *Fundamentals of Geographic Information Systems*, 4th edn, John Wiley & Sons, Ltd, Chichester.
- Dikau, R. and Saurer, H. (1999) *GIS for Earth Surface Systems*, Gebrüder Bornträger Berlin, Stuttgart.
- Foody, G.M. and Atkinson, P.M. (eds) (2002) *Uncertainty in Remote Sensing and GIS*, John Wiley & Sons, Ltd, Chichester.
- Iliffe, J.C. (2000) *Datums and Map Projections for Remote Sensing, GIS and Surveying*, Whittles Publishing, Caithness.
- Kelly, R.E., Drake, N. and Barr, S.L. (eds) (2004) *Spatial Modelling of the Terrestrial Environment*, John Wiley & Sons, Ltd, Chichester.
- Longley, P.A., Goodchild, M.F., Maguire, D.J. and Rhind, D.W. (2005) *Geographic Information Systems and Science*, 2nd edn, John Wiley & Sons, Ltd, Chichester.
- Malcewski, J. (1999) *GIS and Multicriteria Decision Analysis*, John Wiley & Sons, Ltd, Chichester.
- Morain, S. and Lopez Baros, S. (eds) (1996) *Raster Imagery in Geographical Information Systems*, OnWord Press, Albany, NY.
- Petrie, G. and Kennie, T.J.M. (eds) (1990) *Terrain Modelling in Surveying and Civil Engineering*, Whittles Publishing, Caithness.
- Tomlin, C.D. (1990) *Geographic Information Systems and Cartographic Modeling*, Prentice Hall, Upper Saddle River, NJ.
- Wilson, J.P. and Gallant, J.C. (eds) (2000) *Terrain Analysis: Principles and Applications*, John Wiley & Sons, Ltd, Chichester.

## Remote sensing

- Colwell, R.N. (Editor-in-Chief) (1983) *Manual of Remote Sensing – Theory, Instruments and Techniques*, 2nd edn, vol. 1, American Society of Photogrammetry, Falls Church, VA.

Drury, S.A. (2001) *Image Interpretation in Geology*, 3rd edn, Blackwell Science, Cheltenham.

Elachi, C. (1987) *Introduction to the Physics and Techniques of Remote Sensing*, John Wiley & Sons, Inc., New York.

Lilesand, T.M. and Kiefer, R.W. (2000) *Remote Sensing and Image Processing*, 6th edn, John Wiley & Sons, Inc., New York.

Sabins, F.F. (1996) *Remote Sensing: Principles and Interpretation*, 3rd edn, W.H. Freeman, Basingstoke.

## Part One References and further reading

Anderberg, M.R. (1973) *Cluster Analysis for Applications*, Academic Press, New York and London.

Balci, M. and Foroosh, H. (2005) Inferring motion from the rank constraint of the phase matrix. Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2005), 18–23 March, Philadelphia, USA, vol. II, pp. 925–928.

Ball, G.H. (1965) Data analysis in social sciences: what about details? Proceedings of the Fall Joint Computer Conference, pp. 533–539.

Ball, G.H. and Hall, D.J. (1967) A clustering technique for summarizing multivariable data. *Behavioural Science*, **12**, 153–155.

Barron, J.L., Fleet, D.J. and Beauchemin, S.S. (1994) Performance of optical flow techniques. *International Journal of Computer Vision*, **12** (1), 43–77.

Black, M.J. and Jepson, A.D. (1996) Estimating optical flow in segmented images using variable-order parametric models with local deformations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **18** (10), 972–986.

Chavez, P.S. Jr (1989) Extracting spectral contrast in Landsat thematic mapper image data using selective principle component analysis. *Photogrammetric Engineering and Remote Sensing*, **55** (3), 339–348.

Chavez, P.S., Jr, Berlin, G.L. and Sowers, L.B. (1982) Statistical method for selecting Landsat MSS ratios. *Journal of Applied Photographic Engineering*, **8**, 23–30.

Chavez, P.S., Jr, Guptill, S.C. and Bowell, J. (1984) Image processing techniques for Thematic Mapper Data. Proceedings of the 50th Annual ASP–ACSM Symposium, American Society of Photogrammetry, Washington, DC, USA, pp. 728–743.

Cooper, J., Venkatesh, S. and Kitchen, L. (1993) Early jump-out corner detectors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **15**, 823–828.

Corr, D.G. and Whitehouse, S.W. (1996) Automatic change detection in space borne SAR imagery. Proceedings

- of the AGARD Conference – Remote Sensing: A Valuable Source of Information, AGARD-CP-582, October, NATO, Toulouse, France, Paper No. 39.
- Crippen, R.E. (1989) Selection of Landsat TM band and band ratio combinations to maximize lithological information in color composite displays. Proceedings of the 7th Thematic Conference on Remote Sensing for Exploration Geology, 2–6 October, Calgary, Alberta, Canada, 2, pp. 917–919.
- Crist, E.P. and Cicone, R.C. (1984) A physically-based transformation of thematic mapper data – the TM tasseled cap. *IEEE Transactions on Geoscience and Remote Sensing*, **22** (3), 256–263.
- Crosta, A.P., De Souza Filho, C.R., Azevedo, F. and Brodie, C. (2003) Targeting key alteration minerals in epithermal deposits in Patagonia, Argentina using ASTER imagery and principal component analysis. *International Journal of Remote Sensing*, **24** (21), 4233–4240.
- Crosta, A.P. and Moore, J.McM. (1989) Enhancement of Landsat thematic mapper imagery for residual soil mapping in SW Minas Gerais State, Brazil: a prospecting case history in Greenstone Belt terrain. Proceedings of the 7th Thematic Conference on Remote Sensing for Exploration Geology, 2–6 October, Calgary, Alberta, Canada, 2, pp. 1173–1187.
- Crosta, A.P., Sabine, C. and Taranik, J.V. (1998) Hydrothermal alteration mapping at Bodie, California using AVIRIS hyperspectral data. *Remote Sensing of Environment*, **65** (3), 309–319.
- Curlander, J.C. and McDonough, R.N. (1991) *Synthetic Aperture Radar: System and signal processing*, John Wiley & Sons, Inc., New York.
- Devijver, P.A. and Kittler, J. (1982) *Pattern Recognition – A Statistical Approach*, Prentice-Hall International, London.
- Diday, E. and Simon, J.C. (1976) Cluster analysis, in *Digital Pattern Recognition* (ed. K.S. Fu), Springer-Verlag, Berlin, pp. 47–74.
- Dougherty, L., Asmuth, J.C., Blom, A.S. et al. (1999) Validation of an optical flow method for tag displacement estimation. *IEEE Transactions on Medical Imaging*, **18** (4), 359–363.
- Foroosh, H., Zerubia, J.B. and Berthod, M. (2002) Extension of phase correlation to subpixel registration. *IEEE Transactions on Image Processing*, **11** (3), 188–200.
- Förstner, W. (1994) A framework for low level feature extraction. Proceedings of the 3rd European Conference on Computer Vision, Stockholm Sweden, pp. 383–394.
- Fraster, R.S. (1975) Interaction mechanisms within the atmosphere, in *Manual of Remote Sensing*, 1st edn (ed. Robert G. Reeves), American Society of Photogrammetry, Falls Church, VA, pp. 181–233.
- Gabriel, A.K., Goldstein, R.M. and Zebker, H.A. (1989) Mapping small elevation changes over large areas: differential radar interferometry. *Journal of Geophysical Research*, **94** (B7), 9183–9191.
- Gens, R. and van Genderen, J.L. (1996) SAR interferometry – issues, techniques, applications. *International Journal of Remote Sensing*, **17**, 1803–1835.
- Gillespie, A.R., Kahle, A.B. and Walker, R.E. (1986) Color enhancement of highly correlated images I. Decorrelation and HSI contrast stretches. *Remote Sensing of Environment*, **20**, 209–235.
- Goetz, A.F.H., Billingsley, F.C., Gillespie, A.R. et al. (1975) Applications of ERTS images and image processing to regional geologic problems and geologic mapping in northern Arizona. JPL Technical Report 32-1597, Pasadena, CA.
- Goldstein, R.M., Engelhardt, H., Kamb, B. and Frohlich, R.M. (1993) Satellite radar interferometry for monitoring ice sheet motion: application to an Antarctic ice stream. *Science*, **262**, 1525–1634.
- Green, A.A. and Craig, M.D. (1985) Analysis of aircraft spectrometer data with logarithmic residuals. Proceedings of the Airborne Imaging Spectrometer Data Analysis Workshop, JPL Publication 85-41, Pasadena, CA.
- Hagberg, J.O., Ulander, L.M.H. and Askne, J. (1995) Repeat-pass SAR-interferometry over forested terrain. *IEEE Transactions on Geoscience & Remote Sensing*, **33**, 331–340.
- Harris, C. and Stephens, M. (1988) A combined corner and edge detector. Alvey Vision Conference, Manchester, UK, pp. 147–151.
- Heuvelink, G.B.M., Burrough, P.A. and Stein, A. (1989) Propagation of errors in spatial modelling with GIS. *International Journal of Geographical Information Systems*, **3**, 303–322.
- Hoge, W.S. (2003) Subspace identification extension to the phase correlation method. *IEEE Transactions on Medical Imaging*, **22** (2), 277–280.
- Huang, C., Wylie, B., Yang, L. et al. (2002) Derivation of a tasseled cap transformation based on Landsat 7 at-satellite reflectance. *International Journal of Remote Sensing*, **23** (8), 1741–1748.
- Ichoku, C., Karnieli, A., Arkin, Y. et al. (1998) Exploring the utility potential of SAR interferometric coherence images. *International Journal of Remote Sensing*, **19**, 1147–1160.
- Kauth, R.J. and Thomas, G. (1976) The tasseled cap – a graphic description of the spectral-temporal development of agriculture crops as seen by Landsat. Proceedings of the Symposium on Machine Processing of Remotely-sensed Data, Purdue University, West Lafayette, IN, USA, 4B, pp. 41–51.

- Kittler, J. and Pairman, D. (1988) Optimality of reassignment rules in dynamic clustering. *Pattern Recognition*, **21**, 169–174.
- Kruger, S. and Calway, A. (1998) Image registration using multi-resolution frequency domain correlation. Proceedings of the British Machine Vision Conference, 14–17 September, Southampton, UK, pp. 316–325.
- Kruse, F.A., Lefkoff, A.B. and Dietz, J.B. (1993) The spectral image processing system (SIPS) – interaction visualisation and analysis of imaging spectrometer data. *Remote Sensing of Environment*, **44** (Special Issue on AVIRIS, May–June), 145–163.
- Lee, H. and Liu, J.G. (2001) Analysis of topographic decorrelation in SAR interferometry using ratio coherence imagery. *IEEE Transactions on Geoscience and Remote Sensing*, **39** (2), 223–232.
- Liu, J.G. (1990) Hue image colour composition – a simple technique for shadow suppression and spectral enhancement. *International Journal of Remote Sensing*, **11**, 1521–1530.
- Liu, J.G. (1991) Balance contrast enhancement technique and its application in image colour composition. *International Journal of Remote Sensing*, **12**, 2133–2151.
- Liu, J.G. (2000) Smoothing filter based intensity modulation: a spectral preserve image fusion technique for improving spatial details. *International Journal of Remote Sensing*, **21** (18), 3461–3472.
- Liu, J.G., Black, A., Lee, H. et al. (2001) Land surface change detection in a desert area in Algeria using multi-temporal ERS SAR coherence images. *International Journal of Remote Sensing*, **22** (13), 2463–2477.
- Liu, J.G., Capes, R., Haynes, M. et al. (1997a) ERS SAR multi-temporal coherence image as a tool for sand desert study (dune movement, sand encroachment and erosion). Proceedings of the 12th International Conference and Workshop on Applied Geologic Remote Sensing, 17–19 November, Denver, USA, pp. I-478–I-485.
- Liu, J.G. and Haigh, J.D. (1994) A three-dimensional feature space iterative clustering method for multi-spectral image classification. *International Journal of Remote Sensing*, **15** (3), 633–644.
- Liu, J.G., Lee, H. and Pearson, T. (1999) Detection of rapid erosion in SE Spain using ERS SAR interferometric coherence imagery. In *Remote Sensing for Earth Science, Ocean, and Sea Ice Applications* (eds. G. Cecchi, E.T. Engman and E. Zilioli), Proceedings of SPIE, vol. 3868, pp. 525–535.
- Liu, J.G., Mason, P.J., Hilton, F. and Lee, H. (2004) Detection of rapid erosion in SE Spain: a GIS approach based on ERS SAR coherence imagery. *Photogrammetric Engineering and Remote Sensing*, **70** (10), 1179–1185.
- Liu, J.G., Mason, P.J. and Ma, J. (2006) Measurement of the left-lateral displacement of Ms 8.1 Kunlun earthquake on 14th November 2001 using Landsat-7 ETM+ imagery. *International Journal of Remote Sensing*, **27** (10), 1875–1891.
- Liu, J.G. and Moore, J.McM. (1990) Hue image colour composition – a simple technique for shadow suppression and spectral enhancement. *International Journal of Remote Sensing*, **11**, 1521–1530.
- Liu, J.G. and Moore, J.McM. (1996) Direct decorrelation stretch technique for RGB colour composition. *International Journal of Remote Sensing*, **17**, 1005–1018.
- Liu, J.G., Moore, J.McM. and Haigh, J.D. (1997b) Simulated reflectance technique for ATM image enhancement. *International Journal of Remote Sensing*, **18**, 243–255.
- Liu, J.G. and Morgan, G. (2006) FFT selective and adaptive filtering for removal of systematic noise in ETM+ imageodesy images. *IEEE Transactions on Geoscience and Remote Sensing*, **44** (12), 3716–3724.
- Liu, J.G. and Yan, H. (2006) Robust phase correlation methods for sub-pixel feature matching. Proceedings of the 1st Annual Conference of Systems Engineering for Autonomous Systems, Defence Technology Centre, July, Edinburgh, UK, p. A13.
- Liu, J.G. and Yan, H. (2008) Phase correlation pixel-to-pixel image co-registration based on optical flow and median shift propagation. *International Journal of Remote Sensing*, **29** (20), 5943–5956.
- Lucas, B. and Kanade, T. (1981) An iterative image registration technique with application to stereo vision. Proceedings of the 7th International Joint Conference on Artificial Intelligence, 24–28 August, University of British Columbia, Vancouver, BC, Canada, pp. 674–679.
- MacQueen, J. (1967) Some method for classification and analysis of multi-variate observations. Proceedings of the 5th Berkeley Symposium on Mathematics, Statistics and Probability, University of California Press, Berkeley, CA, pp. 281–299.
- Massonnet, D. and Adragna, F. (1993) A full scale validation of radar interferometry with ERS-1: the Landers earthquake. *Earth Observation Quarterly*, **41**, 1–5.
- Massonnet, D., Briole, P. and Arnaud, A. (1995) Deflation of Mount Etna monitored by spaceborne radar interferometry. *Nature*, **375**, 567–570.
- Massonnet, D., Feigl, K., Rossi, M. and Adragna, F. (1994) Radar interferometric mapping of deformation in the year after the Landers earthquake. *Nature*, **369**, 227–230.
- Massonnet, D., Rossi, M., Carmona, C. et al. (1993) The displacement of the Landau earthquake mapped by radar interferometry. *Nature*, **364**, 138–142.

- Olmsted, C. (1993) Alaska SAR Facility Scientific User's Guide. <http://www.asf.alaska.edu/content/reference/SciSARUserGuide.pdf>, 57 (accessed 20 February 2009).
- Reddy, B.S. and Chatterji, B.N. (1996) An FFT-based technique for translation, rotation, and scale-invariant image registration. *IEEE Transactions on Image Processing*, **5** (8), 1266–1271.
- Robinson, N. (1966) *Solar Radiation*, Elsevier, Amsterdam.
- Rosen, P.A., Hensley, S., Zebker, H.A. et al. (1996) Surface deformation and coherence measurements of Kilauea Volcano, Hawaii, from SIR-C radar interferometry. *Journal of Geophysical Research – Planets*, **101** (E10), 23, 109–125.
- Schmid, C., Mohr, R. and Bauckhage, C. (2000) Evaluation of interest point detectors. *International Journal of Computer Vision*, **27** (2), 151–172.
- Sclove, S.L. (1983) Application of conditional population-mixture model to image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **PAMI-5**, 428–433.
- Sheffield, C. (1985) Selecting band combinations from multi-spectral data. *Photogrammetric Engineering and Remote Sensing*, **51** (6), 681–687.
- Shi, Jianbo and Tomasi, C. (1994) Good features to track. IEEE Conference on Computer Vision and Pattern Recognition, June, Seattle, USA, pp. 593–600.
- Smith, A.R. (1978) Colour gamut transform pairs. Proceedings of SIGGRAPH 1978 Conference, 23–25 August, Atlanta, GA, ACM, New York, 3, pp. 12–19.
- Soha, J.M. and Schwartz, A.A. (1978) Multispectral histogram normalisation contrast enhancement. Proceedings of the 5th Canadian Symposium on Remote Sensing, 28–31 August, Victoria, BC, Canada, pp. 86–93.
- Stiller, C. and Konrad, J. (1999) Estimating motion in image sequences. *IEEE Signal Processing Magazine*, **16** (4), 70–91.
- Taylor, M.M. (1973) Principal component colour display of ERTS imagery. The Third Earth Resources Technology Satellite-1 Symposium, 10–14 December, NASA SP-351, pp. 1877–1897.
- Turner, S., Liu, J.G., Cosgrove, J.W. and Mason, P.J. (2006) Envisat ASAR interferometry measurement of earthquake deformation in the Siberian Altai. AGU Western Pacific Geophysics Meeting Conference, 24–27 July, Beijing, China, G21A-03.
- Wang, F., Prinet, V. and Ma, S. (2001) A vector filtering technique for SAR interferometric phase images. Proceedings of Applied Informatics (AI2001), 19–22 February, Innsbruck, Austria, pp. 566–570.
- Zebker, H. and Goldstein, R.M. (1986) Topography mapping from interferometric synthetic aperture radar observations. *Journal of Geophysics Research*, **91**, 4993–4999.
- Zebker, H., Rosen, P., Goldstein, R. et al. (1994a) On the derivation of coseismic displacement fields using differential radar interferometry: the Landers earthquake. *Journal of Geophysics Research*, **99**, 19617–19634.
- Zebker, H., Werner, C., Rosen, P. and Hensley, S. (1994b) Accuracy of topographic maps derived from ERS-1 interferometric radar. *IEEE Transactions on Geosciences and Remote Sensing*, **32**, 823–836.
- Part Two References and further reading**
- Agterberg, F.P. (1974) *Geomathematics: Mathematical Background and Geo-Science Applications*, Elsevier, Amsterdam.
- Agumya, A. and Hunter, G.J. (2002) Responding to the consequences of uncertainty in geographical data. *International Journal of Geographical Information Science*, **16** (5), 405–417.
- Ambroii, T. and Turk, G. (2003) Prediction of subsidence due to underground mining by artificial neural networks. *Computers & Geosciences*, **29** (5), 627–637.
- An, P., Moon, W.M. and Rencz, A. (1991) Application of fuzzy set theory for integration of geological, geophysical and remote sensing data. *Canadian Journal of Exploration Geophysics*, **27**, 1–11.
- Band, L.E. (1986) Topographic partition of watersheds with digital elevation models. *Water Resources Research*, **22** (1), 15–24.
- Burrough, P.A. (1981) Fractal dimensions of landscapes and other environmental data. *Nature*, **294**, 240–242.
- Burrough, P.A. and Frank, A.U. (eds) (1996) *Geographic Objects with Indeterminate Boundaries*, Taylor & Francis, Basingstoke.
- Carter, J. (1992) The effect of data precision on the calculation of slope and aspect using gridded DEMs. *Cartographica*, **29** (1), 22–34.
- Chiles, J.-P. and Delfiner, P. (1999) *Geostatistics: Modeling Spatial Uncertainty*, John Wiley and Sons, Inc., New York.
- Chung, C.F. and Fabbri, A.G. (2003) Validation of spatial prediction models for landslide hazard mapping. *Natural Hazards*, **30**, 451–472.
- Chung, C.F., Fabbri, A.G. and Chi, K.H. (2002) A strategy for sustainable development of nonrenewable resources using spatial prediction models, in *Geoenvironmental Deposit Models for Resource Exploitation and Environmental Security* (eds A.G. Fabbri, G. Gáál and R.B. McCammon), Kluwer Academic, Dordrecht.
- Chung, C.-J. (2006) Using likelihood ratio functions for modeling the conditional probability of occurrence of future landslides for risk assessment. *Computers & Geosciences*, **32** (8), 1052–1068.

- Chung, C.-J. and Fabbri, A.G. (2008) Predicting future landslides for risk analysis – models and cross-validation of their results. *Geomorphology*, **94**, 438–452.
- Chung, C.-J. and Keating, P.B. (2002) Mineral potential evaluation based on airborne geophysical data. *Exploration Geophysics*, **33**, 28–34.
- Cicerone, S. and Clementini, E. (2003) Efficient estimation of qualitative topological relations based on the weighted walkthroughs model. *GeoInformatica*, **7** (2), 211–227.
- Clark, I. (1979) *Practical Geostatistics*, Applied Sciences Publishers, London.
- Corripio, J.G. (2003) Vectorial algebra algorithms for calculating terrain parameters from DEMs and solar radiation modelling in mountainous terrain. *International Journal of Geographical Information Science*, **17** (1), 1–23.
- Cressie, N.A.C. (1993) *Statistics for Spatial Data*, John Wiley and Sons, Inc., New York.
- Danneels, G., Havenith, H.B., Caceres, F. et al. (2007) Filtering of ASTER DEMs using mathematical morphology. *Special Publication of the Geological Society*, In press.
- Dempster, A.P. (1967) Upper and lower probabilities induced by a multi-valued mapping. *Annals of Mathematics Statistics*, **38**, 325–339.
- Dempster, A.P. (1968) A generalization of Bayesian inference. *Journal of the Royal Statistical Society, Series B*, **30**, 205–247.
- Douglas, D.H. and Peucker, T.K. (1973) Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. *Canadian Cartographer*, **10** (2), 112–122.
- Fabbri, A.G. (1984) *Image Processing of Geological Data*, Van Nostrand Reinhold, New York.
- Flacke, W. and Kraus, B. (2005) *Working with Projections and Datum Transformations in ArcGIS*, Points Verlag, Norden Halmstad.
- Gaddy, D.E. (2003) *Introduction to GIS for the Petroleum Industry*, PenWell, Tulsa, OK.
- Goodchild, M.F. (1980) Fractals and the accuracy of geographical measures. *Mathematical Geology*, **12** (2), 85–98.
- Goodchild, M.F. and Gopal, S. (1989) *The Accuracy of Spatial Databases*, Taylor & Francis, New York.
- Gordon, J. and Shortliffe, E.H. (1985) A method for managing evidential reasoning in a hierarchical hypothesis space. *Artificial Intelligence*, **26**, 323–357.
- Hart, J.F. (1954) Central tendency in aerial distributions. *Economic Geography*, **30** (1), 48–59.
- Hart, B.S. and Sagan, J.A. (2007) Curvature for visualization of seismic geomorphology. *Geological Society Special Publications*, **277**, 139–149.
- He, C., Jiang, J., Han, G. and Chen, J. (2004) An algorithm for building full topology. Proceedings of the XXth ISPRS Congress, 12–23 July, Istanbul, Turkey.
- Hunter, G.J. and Goodchild, M.F. (1997) Modeling the uncertainty of slope and aspect estimates derived from spatial databases. *Geographical Analysis*, **29** (1), 35–49.
- Isaaks, E.H. and Srivastava, R.M. (1989) *An Introduction to Applied Geostatistics*, Oxford University Press, Oxford.
- Jenson, S. and Domingue, J. (1988) Extracting topographic structure from digital elevation data for geographic information system analysis. *Photogrammetric Engineering and Remote Sensing*, **54** (11), 1593–1600.
- Jiang, H. and Eastman, J.R. (2001) Application of fuzzy measures in multi-criteria evaluation in GIS. *International Journal of Geographical Information Science*, **14** (2), 173–184.
- Jones, K.H. (1998) A comparison of algorithms used to compute hill slope as a property of the DEM. *Computer and Geosciences*, **24** (4), 315–323.
- Keeney, R.L. and Raiffa, H. (1976) *Decisions with Multiple Objectives: Preferences and Value Tradeoffs*, John Wiley & Sons, Inc., New York.
- Knox-Robinson, C.M. (2000) Vectorial fuzzy logic: a novel technique for enhanced mineral prospectivity mapping, with reference to the orogenic gold mineralisation potential of the Kalgoorlie Terrane, Western Australia. *Australian Journal of Earth Sciences*, **47** (5), 929–941.
- Knox-Robinson, C.M. and Wyborn, L.A.I. (1997) Towards a holistic exploration strategy: using Geographic Information System as a tool to enhance exploration. *Australian Journal of Earth Sciences*, **44**, 453–463.
- Krige, Daniel G. (1951) A statistical approach to some basic mine valuation problems on the Witwatersrand. *Journal of the Chemistry, Metallurgy and Mining Society of South Africa*, **52** (6), 119–139.
- Lam, N. and De Cola, L. (1993) *Fractals in Geography*, Prentice Hall, Englewood Cliffs, NJ.
- Liu, J.G., Mason, P.J., Clerici, N. et al. (2004a) Landslide hazard assessment in the Three Gorges area of the Yangtze River using ASTER imagery: Zigui-Badong. *Geomorphology*, **61**, 171–187.
- Liu, J.G., Mason, P.J., Hilton, F. and Lee, H. (2004b) Detection of rapid erosion in SE Spain: a GIS approach based on ERS SAR coherence imagery. *Photogrammetric Engineering and Remote Sensing*, **70** (10), 1179–1185.
- Lodwick, W.A. (1990) Analysis of structure in fuzzy linear programming. *Fuzzy Sets and Systems*, **38** (1), 15–26.

- Malcewski, J. (2006) GIS-based multicriteria decision analysis: a survey of the literature. *International Journal of Geographical Information Science*, **20** (7), 703–726.
- Maling, D.H. (1992) *Coordinate Systems and Map Projections*, 2nd edn, Pergamon, Oxford.
- Mason, P.J. (1998) Landslide hazard assessment using remote sensing and GIS techniques. PhD Thesis. University of London.
- Mason, P.J. and Rosenbaum, M.S. (2002) Geohazard mapping for predicting landslides: the Langhe Hills in Piemonte, NW Italy. *Quarterly Journal of Engineering Geology & Hydrology*, **35**, 317–326.
- Matheron, G. (1963) Principles of geostatistics. *Economic Geology*, **58**, 1246–1266.
- Matheron, G. (1975) *Random sets and integral geometry*, John Wiley & Sons, Inc., New York.
- Meisels, A., Raizman, S. and Karnieli, A. (1995) Skeletonizing a DEM into a drainage network. *Computers and Geosciences*, **21** (1), 187–196.
- Mineter, M.J. (2003) A software framework to create vector-topology in parallel GIS operations. *International Journal of GIS*, **17** (3), 203–222.
- Minkowski, H. (1911) *Gesammelte Abhandlungen* (ed. D. Hilbert), Leipzig.
- Openshaw, S., Charlton, M. and Carver, S. (1991) Error propagation: a Monte Carlo simulation, in *Handling Geography Information: Methodology and Potential Applications* (eds I. Masser and M. Blakemore), John Wiley & Sons, Inc., New York, pp. 102–114.
- Parker, J.R. (1997) *Algorithms for Image Processing and Computer Vision*, John Wiley & Sons, Inc., New York.
- Reichenbach, P., Pike, R.J., Acevedo, W. and Mark, R.K. (1993) A new landform map of Italy in computer-shaded relief. *Bollettino de Geodesia e Scienze Affini*, **52** (1), 21–44.
- Ross, T.J. (1995) *Fuzzy Logic with Engineering Applications*, John Wiley & Sons, Ltd, Chichester.
- Saaty, T.L. (1980) *The Analytical Heirarchy Process*, McGraw-Hill, New York.
- Saaty, T.L. (1990) *The Analytic Hierarchy Process: Planning, Priority Setting, Resource Allocation*, 2nd edn, with new material, RWS Publications, Pittsburgh, PA.
- Sentz, K. and Ferson, S. (2002) Combination of evidence in Dempster-Shafer theory, SANDIA Tech. Report, 96.
- Serra, J. (1982) *Image Analysis and Mathematical Morphology*, Academic Press, London and New York.
- Shafer, G. (1976) *A Mathematical Theory of Evidence*, Kluwer Academic, Boston, MA.
- Skidmore, A.K. (1989) A comparison of techniques for calculating gradient and aspect from a gridded digital elevation model. *International Journal of Geographical Information Systems*, **3**, 323–334.
- Snyder, J.P. (1997) *Map Projections – A Working Manual*, US Geological Survey Professional Paper 1395, USGS, Washington, DC.
- Snyder, J.P. and Voxland, P.M. (1989) An Album of Map Projections, US Geological Survey Professional Paper 1453, USGS, Washington, DC.
- Soille, P. (ed.) (2003) *Morphological Image Analysis. Principles and Applications*, 2nd edn, Springer-Verlag, Berlin.
- Stein, A., van der Meer, F.D. and Gorte, B.G.H. (eds) (1999) *Remote Sensing and Digital Image Processing*, vol. 1, Kluwer Academic, Dordrecht.
- Tangestani, M.H. and Moore, F. (2001) Porphyry copper potential mapping using the weights-of-evidence model in a GIS, northern Shahr-e-Babak, Iran. *Australian Journal of Earth Sciences*, **48**, 695–701.
- Tobler, W. (1970) A computer movie simulating urban growth in Detroit region. *Economic Geography*, **46**, 234–240.
- Unwin, D.J. (1989) Fractals in the geosciences. *Computers and Geosciences*, **15** (2), 163–165.
- Varnes, D. (1984) Landslide hazard zonation: a review of principles and practice. Commission on Landslides of the International Association of Engineering Geology, United Nations Educational Social and Cultural Organisation, Natural Hazards, No. 3.
- Vincent, L. (1993) Morphological greyscale reconstruction in image analysis: applications and efficient algorithms. *IEEE Transactions on Geoscience and Remote Sensing*, **2** (2), 176–201.
- Wadge, G. (1998) The potential of GIS modeling of gravity flows and slope instabilities. *International Journal of Geographical Information Science*, **2** (2), 143–152.
- Wang, J.J., Robinson, G.J. and White, K. (1996) A fast solution to local viewshed computation using grid-based digital elevation models. *Photogrammetric Engineering and Remote Sensing*, **62**, 1157–1164.
- Watson, D.F. (1992) *Contouring: A Guide to the Analysis and Display of Spatial Data*, Pergamon, Oxford.
- Welch, R., Jordan, T., Lang, H. and Murakami, H. (1998) ASTER as a source for topographic data in the late 1990's. *IEEE Transactions on Geoscience and Remote Sensing*, **36**, 1282–1289.
- Whitney, H. (1932) Congruent graphs and the connectivity of graphs. *American Journal of Mathematics*, **54**, 150–168.
- Wiechel, H. (1878) Theorie und Darstellung der Beleuchtung von nicht gesetzmässig gebildeten Flächen mit Rücksicht auf die Bergzeichnung. *Civilingenieur*, 24.
- Wolf, P.R. and Dewitt, B.A. (2000) *Elements of Photogrammetry (with Applications in GIS)*, 3rd edn, McGraw-Hill, New York.

- Wood, J. (1996) The geomorphological characterisation of digital elevation models, PhD Thesis. University of Leicester.
- Wright, D.F. and Bonham-Carter, G.F. (1996) VHMS favourability mapping with GIS-based integration models, Chisel Lake–Anderson Lake area, in *EX-TECH I: A Multidisciplinary Approach to Massive Sulfide Research in the Rusty Lake–Snow Lake Greenstone Belts* (eds G.F. Bonham-Carter et al.), Geological Survey of Canada, Manitoba, Bulletin 426, pp. 339–376.
- Wu, Y. (2000) R2V conversion: why and how? *GeoInformatics*, **3** (6), 28–31.
- Xia, Y. and Ho, A.T.S. (2000) 3D vector topology model in the visualization system. Proceedings of IGARSS 2000, the IEEE International Symposium on Geoscience and Remote Sensing, Honolulu, USA, 7, pp. 3000–3002.
- Yang, Q., Snyder, J. and Tobler, W. (2000) *Map Projection Transformation: Principles and Applications*, Taylor & Francis, Abingdon.
- Zadeh, L.A. (1965) Fuzzy sets. *IEEE Information and Control*, **8**, 338–353.
- Zevenbergen, L.W. and Thorne, C.R. (1987) Quantitative analysis of land surface topography. *Earth Surface Processes and Landforms*, **12**, 47–56.
- Ziadat, F.M. (2007) Effect of contour intervals and grid cell size on the accuracy of DEMs and slope derivatives. *Transactions in GIS*, **11** (1), 67–81.
- Zebker, H. and Villasenor, J. (1992) Decorrelation in interferometric radar echoes. *IEEE Transactions on Geoscience and Remote Sensing*, **30** (5), 950–959.

# Index

- 2D 3  
2D fitting 113  
3D 3  
3D vector features 216, 436 (Xia)  
3D-FSIC 93, 432 (Liu)  
3D Index 89, 431 (Crippen)
- Achromatic factor 62  
Across-track mechanical scanner 414  
Active sensor 4, 409  
Adaptive filtering 48  
Adaptive median filter 42  
Additive RGB Colour Composite System 6  
Advanced spatial analysis 241  
Airborne Thematic Mapper 287  
Albedo 30  
Algebraic combination 248  
Algebraic Operations 21  
Algorithm  $\alpha$  92–93  
Algorithm  $\beta$  93  
Along-track push-broom scanner 415  
ALOS 213  
Analytical Hierarchy Process 252  
Analytical method 238  
ANDORness 254  
Arc-node structure 150–151, 434 (Cicerone), 435 (Mineter)  
Area definition 153  
Artificial Illumination 220  
Aspect 220, 433 (Carter), 434 (Hunter), 435 (Skidmore)
- ASST 94  
ASTER 212, 288, 411, 412  
ATM 288, 292  
Atmospheric correction 267  
Attributes 142  
Autocorrelation 112  
Autocorrelation matrix 112, 433 (Schmid)  
AVIRIS 418  
Azimuth 220, 433 (Carter), 434 (Hunter)
- B/W 4  
Balance Contrast Enhancement Technique 16, 432 (Liu)  
Band selection technique 88  
Bands 3  
Bare Earth Model 212  
Baseline 122  
Bayes' Theorum 249  
Bayesian probability 381  
BCET 16, 432 (Liu)  
Belief 251, 435 (Senz), 436 (Zadeh)  
Belief Interval 251  
Bilinear interpolation 110  
Binary operations 184  
Boolean logical combination 248  
Boolean operators 180  
Boundary inclusion 159  
Box classifier 97  
Box filters 37  
Brovey transform 73  
Buffer 191  
Butterworth band-reject filter 53

- C2D-FT 39  
 C2D-IFT 39  
 Cartesian coordinates 166  
 CCD 413  
 Central Meridian 170, 172  
 Channels 3  
 Characteristic equation 79  
 Characteristic polynomial 80  
 Classification 91, *Chapter 8*  
 Classification accuracy assessment 100  
 Clean columns filter 44  
 Clean lines filter 44  
 Clean pixels filter 44  
 Clipping 18, 193  
 Closing 187  
 Clustering 91, 430 (Anderberg), 431 (Devijver, Didy)  
 Coherence 128, 360  
 Co-Kriging 208  
 Colatitude 168  
 Combinatorial operators 180  
 Complementary colour 5  
 Condition independence 249  
 Conditional probability 249  
 Conditional processing 179  
 Conditional smoothing filters 43  
 Conformal 169  
 Confusion matrix 100  
 Conic 168  
 Conic projections 169  
 Conical scan 419  
 Connectivity 150  
 Consistency index 246  
 Contiguity 153  
 Contours 216  
 Contrast enhancement See: Chapter 2  
 Convolution 39  
 Convolution Theorem 39  
 Coordinate 163  
 Coordinate systems 166  
 Co-registration 108, 113  
 Corner detectors 112, 430 (Cooper), 431  
 (Förstner, Harris)  
 Corner points 112  
 Cost pathways 187  
 Cost-weighted analysis 185  
 Cost-weighted direction 187  
 Cost-weighted distance 185, 187  
 Covariance matrix 79  
 Coverage 152  
 Cracking 194  
 Criterion uncertainty 233, 235  
 Cross-correlation 208  
 Cross-validating 238  
 Cumulative distribution function 14  
 Curvature 222, 434 (Hart), 435 (Mason),  
 436 (Wood)  
 Cut and fill analysis 228  
 Cylindrical 168, 169  
 Cylindrical Equal Area 170  
 Cylindrical Equidistant 170  
 D2D-FT 39  
 D2D-IFT 39  
 Data capture 274  
 Data conversion 157  
 Data Models 141  
 Data preparation 282  
 Data quantisation 143  
 Data-driven (empirical) 242  
 Data-driven (neural-network) 243, 433  
 (Ambrogi)  
 Datum 163, 166  
 DDS 61  
 Decision rule uncertainty 234, 236  
 Decision support 231, 434 (Dempster)  
 Decorrelation stretch 59, 61, 85  
 Deformation model 108  
 Delaunay triangulation 156  
 DEM 211, 435 (Welch)  
 Dempster-Shafer theory 251, 435 (Sentz)  
 436 (Zadeh)  
 Deterministic 202  
 Deterministic Decision Making 232  
 Deterministic interpolators 202  
 Difference image 22  
 Differential InSAR 125  
 Digital Elevation Model See: DEM  
 Digital image 3  
 Digital image colour display 5  
 Digital image processing 3  
 Digital Surface Model 212  
 Digital Terrain Model 212  
 Dilation 187  
 Dimensionality and colour confusion reduction 82,  
 430 (Chavez)  
 DInSAR 125, 431 (Gabriel, Goldstein), 432  
 (Massonnet), 433 (Rosen, Zebker)  
 Direct decorrelation stretch See: DDS  
 Disbelief 251  
 Discrete form of convolution 40  
 Disjunctive kriging 208  
 Dissection density 341  
 Dissimilarity functions 97  
 Dissolve 193  
 DN 3  
 Doppler frequency shift 420, 421

- Drainage networks 225, 433 (Band), 434 (Jenson), 435 (Meisels), 436 (Wood)
- Dual Independent Map Encoding (DIME) 150
- Eccentricity 164
- ECW 144
- Edge sharpening filters 47
- Edge-preserve smoothing filter 42, 43
- Effect of resolution 146
- Eigenvalues 79
- Eigenvector 80
- Ellipsoid of rotation 164
- EMR 409
- Enhanced Compressed Wavelet 144
- Enhanced Thematic Mapper Plus 410
- Epithermal gold 287
- Equal area 169
- Equidistant 169
- Equipotential surface 165
- Erosion 187
- Error Assessment 235
- Error propagation 235, 435 (Openshaw)
- ERS 422
- ETM 410
- Euclidean OMPR 98
- Evaluation criteria 243
- Events 155
- Expansion 187
- Exponential contrast enhancement 14
- Extended neighbourhood 191
- Extracting surface parameters 220
- Extrusion 219
- Factor 241, 244
- False colour composite 6
- False colour composition 6, 7
- False Northing & Easting 172
- Favourability 236, 242
- Feature matching 113, 430 (Barron)
- Feature oriented PC selection 81, 82
- Feature space partition 92
- FFT 38
- FFT adaptive filtering 50, 51
- FFT selective filtering 49
- Filtering 37, Chapter 4
- Flattening 164
- Flow accumulation 225
- Flow direction 226
- Focal operations 191
- Fourier rotation property 114
- Fourier scale property 114
- Fourier shift property 113, 114
- Fourier Transform 39
- FPCS 81, 431(Crosta)
- FPCS spectral contrast mapping 84
- Fractals 212, 433 (Band, Burrough), 434 (Goodchild, Lam), 435 (Unwin)
- Fringe smoothing filter 132
- Fuzzy algebraic product 255
- Fuzzy algebraic sum 255
- Fuzzy Confidence 256
- Fuzzy Decision Making 232
- fuzzy gamma function 255
- Fuzzy geohazard mapping 350
- Fuzzy logic 254, 433 (An)
- Fuzzy measures 248, 434 (Jiang)
- Fuzzy membership 248, 254
- Fuzzy prospectivity 256
- Gaussian filter 41
- Gaussian OMPR 98, 99
- Gaussian Stretch 15
- GCP 108
- Geochemical sample 378
- Geocoding 107, 163
- Geodatabase 152
- Geodetic 164
- Geodetic coordinate 166
- Geodetic datum 166
- GeoEye1 212
- Geoid 165
- Geolocation 163
- Geometric distortions 167
- Geometric mean 338
- Geo-rectification 265
- Geo-referencing 107
- Georegistration 163
- Geostatistics 197, 433 (Chiles), 434 (Hart, Krige), 435 (Matheron, Stein)
- Geotechnical 352
- GeoTIFF 174
- Global polynomial 203
- Gradient 45, 220, 434 (Jones), 435 (Skidmore), 436 (Ziadat)
- Gradient filters 45
- Grey line 6
- Ground elevations 219
- Gypsum enhancement 283
- Hazard 234, 433 (Agumya)
- High pass filters 44, 45
- Hillshading See: Shaded relief
- Histogram equalisation (HE) 14
- Histogram matching 15
- Histogram modification 9
- Histogram 198

- HRGB 63, 432 (Liu)  
 HRV 410  
 Hue 57  
 Hybrid classification 91  
 Hyper-spectral sensors 417
- IDW 201, 206  
 IFOV 413  
 IHS decorrelation stretch See: IHDS  
 IHDS 59, 431 (Gillespie)  
 IHS-RGB transform 58, 59  
 Ikonos 213  
 Image addition 21  
 Image classification See: Chapter 8  
 Image compression 144  
 Image division (ratio) 24  
 Image fusion See: Chapter 6  
 Image geometric deformation 105  
 Image histogram 9  
 Image interpretation 270, 274  
 Image multiplication 22  
 Image processing 3  
 Image scanning classification 92  
 Image subtraction 22  
 Index of three-dimensionality See: 3D Index  
 Index-overlay combination 248  
 Indicator Kriging 208  
 Inexact interpolators 203  
 InSAR See: Chapter 10, 214, 431 (Gens, Hagberg),  
     433 (Turner, Zebker)  
 InSAR coherence 127, 430 (Corr), 431 (Ichoku), 432  
     (Liu), 433 (Rosen)  
 Instant Field Of View See: IFOV  
 Intensity 57  
 Interoperability 139  
 Interpolation 201, 202  
 Intersection 193  
 Interval 142  
 Inverse Distance Weighted See: IDW  
 Inverse Distance Weighted Average 206  
 Iron oxide ratio index 28  
 ISNULL 179  
 ISODATA 93, 430 (Ball)  
 Iterative classification 92
- JPEG2000 144
- K nearest mean filter 42  
 K nearest median filter 43  
 kappa coefficient 101, 239  
 Knowledge-driven 242  
 Kriging 207, 434 (Krige), 435  
     (Matheron)
- L-1B data 108  
 Landsat 411  
 Landslide 334, 350  
 Landslide hazard assessment 354  
 Landslide hazard mapping 347  
 Landslide susceptibility mapping 338  
 Landuse 296, 307  
 Laplacian 45  
 Laplacian filters 46  
 Laser Altimetry 213  
 Latitude 164  
 Layer combination 381  
 Layers 3  
 LCE 11  
 Least cost pathway 185, 187  
 Leave-one-out cross-validation 239  
 LiDAR 213  
 Linear contrast enhancement See: LCE  
 linear referencing 154  
 Linear scale 12  
 Lines 143  
 Local contrast enhancement 48  
 Local mean adjustment 48  
 Local neighbourhood 185  
 Local operations 181  
 Local polynomial 203  
 Logarithmic contrast enhancement 13  
 Logarithmic residual 29, 431 (Green)  
 Logical operators 180  
 Logical processing 179  
 Logical relationships 148, 434 (He)  
 Longitude 164  
 LOOCV 239  
 Low-pass filter 40  
 LUT (Look up table) 9
- Map algebra 178  
 Map composition 270, 278  
 Map projection 167  
 Mathematical Morphology 187, 434 (Danneels), 435  
     (Matheron, Minkowski, Vincent)  
 Maximum likelihood classifier 98  
 Mean filters 41  
 Mean/standard deviation adjustment 12  
 Measurement scales 142  
 Measuring height 165  
 Median filter 42  
 Median shift propagation See: MSP  
 Mercator 170  
 Meridian 164  
 Metadata 139, 174  
 Microwave 4  
 Microwave imaging radiometer 418

- Mineral mapping 287  
 Mineral occurrences 376  
 Mineral prospectivity 371  
 Minkowski subtraction 189  
 Minkowski sum operator 188, 435 (Minkowski)  
 Minkowski summation 188  
 Mode (majority) filter 43  
 Monochromatic display 4  
 Monte Carlo Simulation 237, 435 (Openshaw)  
 Morphological Reconstruction 191, 435 (Vincent)  
 MrSID 144  
 MSP 115  
 MSS 413  
 Multi-attribute decision making 236  
 Multi-criteria analysis See: Chapter 18, 433  
     (Chung), 434 (Dempster, Liu), 435 (Malcewski, Mason, Wadge)  
 Multi-criteria assessment 371  
 Multi-criteria combination methods 248  
 Multi-Criteria Decision Making 236  
 Multi-criteria evaluation (MCE) See: Chapter 18, 356, 434 (Knox-Robinson, Liu), 435 (Wadge)  
 Multi-image point operations 21  
 Multi-look SAR image 121  
 Multi-objective decision making 236, 435  
     (Malcewski)  
 Multi-resolution Seamless Image Database  
     See: MrSID  
 Multi-source data integration 73  
 Multi-spectral sensor 409  
 Multi-variable elimination and characterisation 336  
  
 Nadir view imaging 106  
 N-ary operations 185  
 Natural Barriers 200  
 NCC 111  
 NDVI 27  
 Nearest K neighbours 201, 202  
 Neighbourhood operations 185  
 Neighbourhood processing 37  
 Network connectivity 150  
 Noise cleaning filters 43  
 Normalised Difference Vegetation Index See: NDVI  
 Norminal 142  
 Null data 178  
  
 Oblique light shading 216  
 Oblique view imaging 106  
 OIF 88, 430 (Chavez)  
 OMPR 98, 432 (Kittler, MacQueen)  
 Opening 187, 189  
 Optimal Index Factor See: OIF  
 Optimal multiple point reassignment See: OMPR  
  
 Order weights 253  
 Ordered Weighted Average (OWA) 253, 434 (Jiang)  
 Ordinal 142  
 Ordinary kriging 208  
 Orthometric heights 165  
 Output-to-input mapping 109  
 Overall accuracy 101  
  
 Pairwise comparison 238, 245  
 Pairwise comparison matrix 356  
 Parallel classifier 97  
 Parallels 164  
 Passive microwave sensor 418  
 Passive sensor 4, 413  
 PCA See: Chapter 7  
 PCADS 85, 433 (Soha, Taylor)  
 Phase Correlation 113, 114, 430 (Balci), 431  
     (Foroosh, Hoge), 432 (Liu), 433 (Reddy, Stiller)  
 Photogrammetry 212  
 Piecewise linear stretch (PLS) 12  
 Pixel 3  
 Pixel-to-pixel image co-registration 113, 430 (Black), 431 (Dougherty), 432 (Kruger, Liu, Lucas)  
 Plan (Cross-sectional) curvature 222  
 Planar (azimuthal) projections 172  
 Platform flight coordinate system 105, 106  
 Point to raster 158  
 Point allocation 244  
 Point operation 9  
 Point Spread Function (PSF) 39  
 Polygon to raster 160  
 Polyline to raster 159  
 POLYVRT 150, 434 (Cicerone), 435 (Mineter)  
 Prewitt filters 45  
 PRF 420  
 Primary colours 5  
 Primary operations 181  
 Principal component analysis See: PCA  
 Principal components 80  
 Probabilistic Decision Making 232  
 Probability density function 9  
 Producer's accuracy 101  
 Profile (down-slope) curvature 222  
 Projection 164  
 Projection coordinates 166  
 Projection specification 172  
 Pseudo colour display 7  
  
 Radar polarization 421  
 Rank exponent 245  
 Rank order 245  
 Rank reciprocal 245  
 Rank sum 245

Ranking 245, 435 (Malcewski)  
 Raster data 3, 143  
 Raster to point 160  
 Raster to polygon 161  
 Raster to polyline 160, 434 Douglas  
 Raster to vector conversion 160, 434 (Douglas),  
     436 (Wu)  
 Rasterisation 158  
 Rating 244  
 Ratio images 24, 431 (Goetz)  
 Ratio measurement 142  
 Ratio coherence 129, 432 (Lee)  
 Ratio estimation 245  
 RBV 413  
 Real aperture (SLR) 420  
 Reclassification 182  
 Red edge 27  
 Reflective infrared (IR) 4  
 Regions 154  
 Relational operators 179  
 Remote sensing xv  
 Repeat-pass SAR interferometer 122  
 Result validation 238  
 RGB colour cube 6  
 RGB-IHS transform 57, 58  
 Risk 234, 433 (Agumya)  
 Root mean square errors (RMS) 235  
 Routes 155  
 Rubber sheet warping 108  
  
 SAM 96, 432 (Kruse)  
 Sample size 201  
 SAR 4, 419, 420, 431 (Curlander), 433 (Olmsted)  
 SAR Interferometry See: InSAR  
 SAR intensity image 122  
 Saturation 57, 58  
 Scale factor 168, 172  
 Search radius 201  
 Secant 168  
 Sections 155  
 Selecting coordinate systems 174  
 Selective enhancement 267  
 Selective Principal Component Analysis See: SPCA  
 Semivariogram 199  
 Sensitivity analysis 237, 238, 435 (Malcewski,  
     Openshaw)  
 Sensor 3D status 105, 106  
 Sensors and satellites See: Appendix A  
 SFIM 73, 432 (Liu)  
 Shaded relief 216, 217, 434 (Corripi), 435  
     (Reichenbach, Welch)  
 Shapefile 151  
 Short wave infrared See: SWIR

Shrinking 187  
 Side Looking Radar See: SLR  
 Simple kriging 208  
 Simulated irradiance 29, 30  
 Simulated reflectance 29, 30, 432 (Liu)  
 Simulated thermal emittance 31, 432 (Liu)  
 Single look complex See: SLC  
 Single pass classification 92  
 Single-pass SAR interferometer 122  
 SLC 121  
 Slope 220, 434 (Jones), 435 (Skidmore), 436 (Ziadat)  
 Slope instability 335, 350  
 SLR 419  
 SMMR 419  
 Smoothing filters 40  
 Sobel filters 46  
 Soundshed 227  
 Spaghetti data 148  
 Spatial analysis 235, 241  
 Spatial attributes 178  
 Spatial autocorrelation 198, 435 (Tobler)  
 Spatial data 141  
 Spatial Decision Support System 231  
 Spatial decorrelation 129  
 Spatial modeling 381  
 Spatial relationships 145  
 Spatial variability 145  
 SPCA 82  
 Spectral analysis 267, 283  
 Spectral Angle Mapping See: SAM  
 Spectral contrast mapping 84, 430 (Chavez)  
 Spectral indices 26, 390  
 Spectral preserve image fusion technique 73  
 Spherical coordinates 166  
 Spheroidal coordinates 166  
 Spheroidal heights 165  
 Splines 204  
 SPOT 213, 415  
 Standard false colour composite 7  
 Standardisation 29  
 Stationarity 199  
 Statistical operations 186  
 stochastic 202  
 Stochastic interpolators 207, 434 (Krigie)  
 Strategy 261  
 Structural enhancement 313  
 Suitability 236  
 Supervised classification 96  
 Supervised enhancement 26  
 Surface topology 225  
 Surfaces 143, 147, 155, 211  
 SVD 113  
 SWIR 4

- Target area identification 388  
Targeted spectral image processing 377  
Tasselled cap transformation 86, 431  
    (Crist, Huang, Kauth)  
Temporal decorrelation 130  
Thematic Mapper See: TM  
Thematic mapping 138, 271  
Thermal infrared 4  
Thermal infrared sensors 416  
Thermal sensing 416  
Threshold uncertainty 233, 236  
Tie-point 166  
TIN 155  
TM 410  
TM clay mineral ratio index 29  
Topographic information extraction 339  
Topology 148  
Total cost 187  
TRADEOFF 254  
Transverse Mercator 170  
Trends 200  
Triangulated Irregular Network 155, 214  
Tristimulus Colour Theory 5  
TRMM 419  
True colour composite 6, 7  
  
Unary operations 182  
Uncertainty 232, 435 (Malcewski)  
Union 194  
Universal kriging 208  
Universal Transverse Mercator See: UTM  
Unsupervised classification 92  
User's accuracy 101  
Uses of Surface Data 215  
UTM 172  
  
Variograms 199  
Vector data 147, 434 (Cicerone), 435 (Mineter)  
Vector equivalents to raster map algebra 192  
Vector projection 216  
Vector surfaces 214  
Vector to raster conversion 158  
Vectorial fuzzy modeling 256, 434 (Knox-Robinson)  
Vectorisation 160, 434 (Douglas)  
Vertex Dictionary 149  
Vertical exaggeration 219  
Viewshed 226, 435 (Wang)  
Virtual Reality Modelling Language 215  
Visible-photographic infrared 4  
Visualising in three dimensions 218  
Visualising in two dimensions 216  
Visualising surfaces 215  
VNIR 4  
Voronoi polygons 157, 435 (Whitney)  
VRML 215  
Vulnerability 234  
  
Water resource exploration 386  
Watersheds 225, 433 (Band), 434 (Jenson), 435  
    (Meisels) 436 (Wood)  
Wavelet transforms 144  
Weighted factors in Linear Combination See: WLC  
Weighted mean filters 41  
Weighted sum vector (WSV) 246  
Weighting coefficients 244  
Weights of evidence modelling 249, 434 (Chung), 435  
    (Dempster)  
WGS84 173  
WLC 252  
World file 174  
World Geodetic System 1984 See: WGS84  
Worldview-1 213  
  
X3D 215  
  
Zonal operations 192